Probabilistic and Bayesian Machine Learning

Lecture 1: Introduction to Probabilistic Modelling

Yee Whye Teh
ywteh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit
University College London

Why a probabilistic approach?

- Many machine learning problems can be expressed as latent variable problems.
- Given some data, solution can be obtained by inferring the values of unobserved, latent variables.
- There is much uncertainty in the world:
  - Noise in observations.
  - Intrinsic stochasticity.
  - Effects that are complex, unknown, and/or not understood.
  - Our own state of belief being not certain.
- Probability theory gives coherent and simple way to reason about uncertainty.
- Probabilistic modelling gives
  - powerful language to express our knowledge about the world.
  - powerful computational framework for inference and learning about the world.

Basic Rules of Probability

Basic Rules of Probability

Probabilities are non-negative $P(x) \geq 0 \; \forall x$.

Basic Rules of Probability

Probabilities are non-negative $P(x) \geq 0 \; \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(x) = 1$ for distributions if x is a discrete variable and $\int_{-\infty}^{+\infty} p(x) dx = 1$ for probability densities over continuous variables

Warning: I will not be obsessively careful in my use of p and P for probability density and probability distribution. Should be obvious from context.

Basic Rules of Probability

Probabilities are non-negative $P(x) \geq 0 \; \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(x) = 1$ for distributions if x is a discrete variable and $\int_{-\infty}^{+\infty} p(x) dx = 1$ for probability densities over continuous variables

The joint probability of x and y is: $P(x, y)$.

Warning: I will not be obsessively careful in my use of p and P for probability density and probability distribution. Should be obvious from context.

Basic Rules of Probability

Probabilities are non-negative $P(x) \geq 0 \; \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(x) = 1$ for distributions if x is a discrete variable and $\int_{-\infty}^{+\infty} p(x) dx = 1$ for probability densities over continuous variables

The joint probability of x and y is: $P(x, y)$.

The marginal probability of x is: $P(x) = \sum_y P(x, y)$, assuming y is discrete.

Warning: I will not be obsessively careful in my use of p and P for probability density and probability distribution. Should be obvious from context.

Basic Rules of Probability

Probabilities are non-negative $P(x) \geq 0 \; \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(x) = 1$ for distributions if x is a discrete variable and $\int_{-\infty}^{+\infty} p(x)dx = 1$ for probability densities over continuous variables

The joint probability of x and y is: $P(x, y)$.

The marginal probability of x is: $P(x) = \sum_y P(x, y)$, assuming y is discrete.

The conditional probability of x given y is: $P(x|y) = P(x, y)/P(y)$.

Warning: I will not be obsessively careful in my use of p and P for probability density and probability distribution. Should be obvious from context.

Basic Rules of Probability

Probabilities are non-negative $P(x) \geq 0 \; \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(x) = 1$ for distributions if x is a discrete variable and $\int_{-\infty}^{+\infty} p(x) dx = 1$ for probability densities over continuous variables

The joint probability of x and y is: $P(x, y)$.

The marginal probability of x is: $P(x) = \sum_y P(x, y)$, assuming y is discrete.

The conditional probability of x given y is: $P(x|y) = P(x, y)/P(y)$.

Independent random variables: $X \perp\!\!\!\perp Y$ means $P(x, y) = P(x)P(y)$.

Warning: I will not be obsessively careful in my use of p and P for probability density and probability distribution. Should be obvious from context.

Basic Rules of Probability

Probabilities are non-negative $P(x) \geq 0 \; \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(x) = 1$ for distributions if x is a discrete variable and $\int_{-\infty}^{+\infty} p(x) dx = 1$ for probability densities over continuous variables

The joint probability of x and y is: $P(x, y)$.

The marginal probability of x is: $P(x) = \sum_y P(x, y)$, assuming y is discrete.

The conditional probability of x given y is: $P(x|y) = P(x, y)/P(y)$.

Independent random variables: $X \perp\!\!\!\perp Y$ means $P(x, y) = P(x)P(y)$.

Conditional independence: $X \perp\!\!\!\perp Y | Z$ (X conditionally independent of Y given Z) means $P(x, y|z) = P(x|z)P(y|z)$ and $P(x|y, z) = P(x|z)$.

Warning: I will not be obsessively careful in my use of p and P for probability density and probability distribution. Should be obvious from context.

Basic Rules of Probability

Probabilities are non-negative $P(x) \geq 0 \; \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(x) = 1$ for distributions if x is a discrete variable and $\int_{-\infty}^{+\infty} p(x) dx = 1$ for probability densities over continuous variables

The joint probability of x and y is: $P(x, y)$.

The marginal probability of x is: $P(x) = \sum_y P(x, y)$, assuming y is discrete.

The conditional probability of x given y is: $P(x|y) = P(x, y)/P(y)$.

Independent random variables: $X \perp\!\!\!\perp Y$ means $P(x, y) = P(x)P(y)$.

Conditional independence: $X \perp\!\!\!\perp Y | Z$ (X conditionally independent of Y given Z) means $P(x, y|z) = P(x|z)P(y|z)$ and $P(x|y, z) = P(x|z)$.

Bayes Rule:

$$P(x, y) = P(x)P(y|x) = P(y)P(x|y) \quad \Rightarrow \quad \boxed{P(y|x) = \frac{P(x|y)P(y)}{P(x)}}$$

Warning: I will not be obsessively careful in my use of p and P for probability density and probability distribution. Should be obvious from context.

Probability, Information and Entropy

Information is the reduction of uncertainty. How do we measure uncertainty?

Probability, Information and Entropy

Information is the reduction of uncertainty. How do we measure uncertainty?

Some axioms (informal):

- if something is certain its uncertainty = 0
- uncertainty should be maximum if all choices are equally probable
- uncertainty (information) should add for independent sources

Probability, Information and Entropy

Information is the reduction of uncertainty. How do we measure uncertainty?

Some axioms (informal):

- if something is certain its uncertainty = 0
- uncertainty should be maximum if all choices are equally probable
- uncertainty (information) should add for independent sources

This leads to a discrete random variable X having uncertainty equal to the entropy function:

$$H(X) = -\sum_{X=x} P(X = x) \, \log P(X = x)$$

measured in bits (**bi**nary digi**ts**) if the base 2 logarithm is used or nats (**na**tural digi**ts**) if the natural (base e) logarithm is used.

Probability, Information and Entropy

- Surprise (for event $X = x$): $-\log P(X = x)$

Probability, Information and Entropy

- Surprise (for event $X = x$): $-\log P(X = x)$
- Entropy = average surpise: $H(X) = -\sum_{X=x} P(X = x)\ \log P(X = x)$

Probability, Information and Entropy

- Surprise (for event $X = x$): $-\log P(X = x)$
- Entropy = average surpise: $H(X) = -\sum_{X=x} P(X = x)\ \log P(X = x)$
- Conditional entropy

$$H(X|Y) = -\sum_{x}\sum_{y} P(x, y) \log P(x|y)$$

Probability, Information and Entropy

- Surprise (for event $X = x$): $-\log P(X = x)$
- Entropy = average surpise: $H(X) = -\sum_{X=x} P(X = x)\ \log P(X = x)$
- Conditional entropy

$$H(X|Y) = -\sum_x \sum_y P(x, y) \log P(x|y)$$

- Mutual information

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

Probability, Information and Entropy

- Surprise (for event $X = x$): $-\log P(X = x)$
- Entropy = average surpise: $H(X) = -\sum_{X=x} P(X = x)\ \log P(X = x)$
- Conditional entropy

$$H(X|Y) = -\sum_x \sum_y P(x, y) \log P(x|y)$$

- Mutual information

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y)$$

- Kullback-Leibler divergence (relative entropy)

$$KL(P(X)\|Q(X)) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Probability, Information and Entropy

- Surprise (for event $X = x$): $-\log P(X = x)$
- Entropy = average surpise: $H(X) = -\sum_{X=x} P(X = x) \ \log P(X = x)$
- Conditional entropy

$$H(X|Y) = -\sum_x \sum_y P(x, y) \log P(x|y)$$

- Mutual information

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

- Kullback-Leibler divergence (relative entropy)

$$KL(P(X)\|Q(X)) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Relation between mutual information and KL: $I(X;Y) = KL(P(X,Y)\|P(X)P(Y))$

Probabilistic Models and Inference

Describe the world using probabilistic models.

$$P(X, Y|\theta)$$

X: observations, measurements, sensory input about the world.
Y: Unobserved variables.
θ: Parameters of our model.

Inference: Given $X = x$, apply Bayes' Rule to compute posterior distribution over unobserved variables of interest Y:

$$P(Y|x, \theta) = \frac{P(x, Y|\theta)}{P(x|\theta)}$$

Maximum a posteriori: $y^{\text{MAP}} = \operatorname*{argmax}_y P(y|x, \theta).$

Mean: $y^{\text{mean}} = E_{P(Y|x,\theta)}[Y].$

Minimize Loss: $y^{\text{Loss}} = \operatorname*{argmin}_y E_{P(Y|x,\theta)}[\text{Loss}(Y)].$

Probabilistic Models and Learning

It is typically relatively easy to specify by hand high level structural information about $P(X, Y|\theta)$ from prior knowledge.

Much harder to specify exact parameters θ describing the joint distribution.

- (Unsupervised) Learning: Given examples of $x_1, x_2, \ldots$ find parameters θ that "best explains" examples. Typically by maximum likelihood:

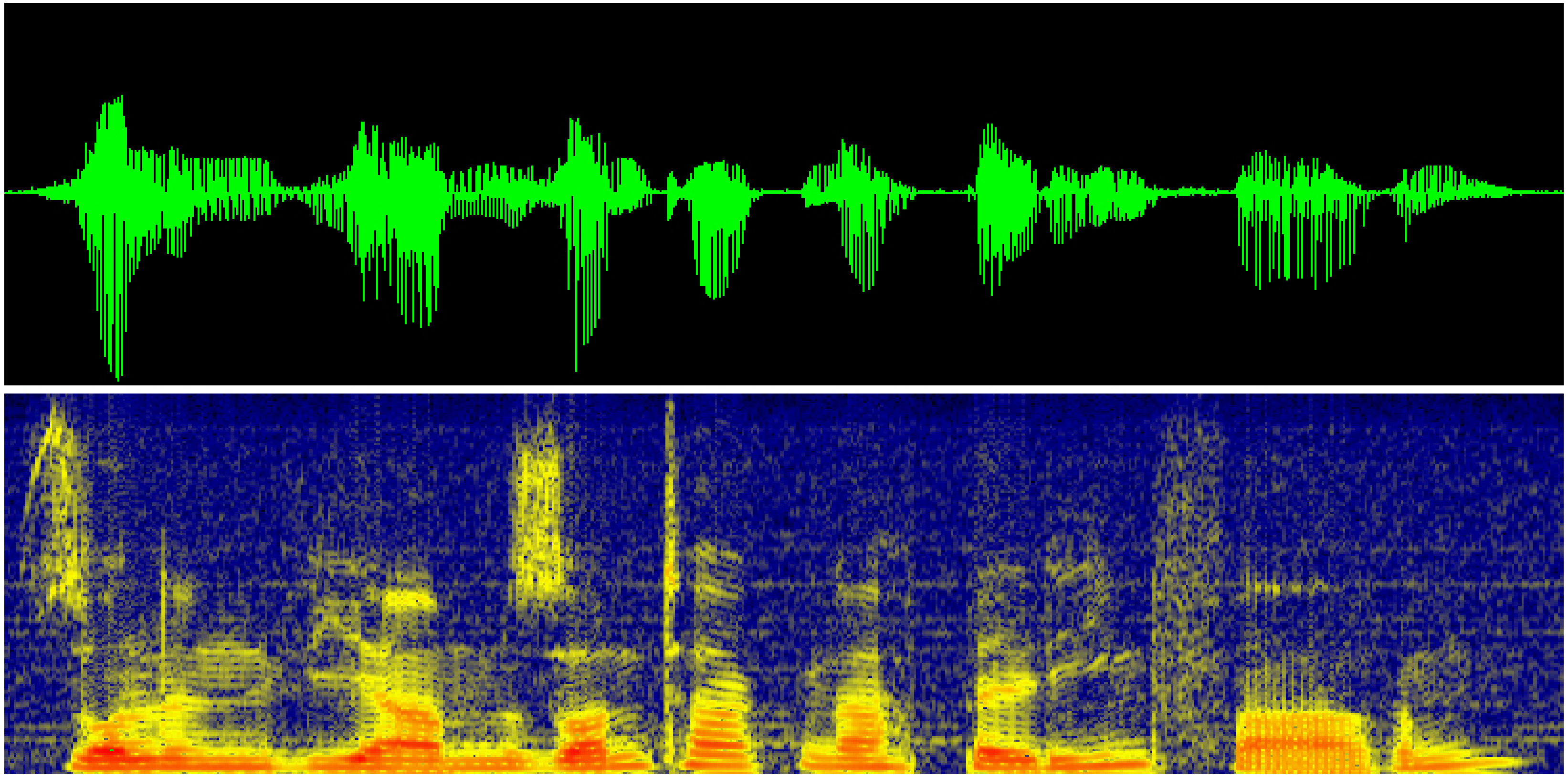
  $$\theta^{ML} = \operatorname*{argmax}_{\theta} P(x_1, x_2, \ldots |\theta)$$

  Alternatives: maximum a posteriori learning, Bayesian learning.

- (Supervised) Learning: Given $y_1, y_2, \ldots$ as well.

Often models will come in a series $\mathcal{M}_1, \mathcal{M}_2, \ldots$ of increasing complexity. Each gives a joint distribution $P(X, Y|\theta_i, \mathcal{M}_i)$. We would like to select one of the right size to avoid over-fitting or under-fitting.

Speech Recognition



$$Y = \text{word sequence}, \qquad P(Y|\theta) = \text{language model},$$
$$X = \text{acoustic signal}, \qquad P(X|Y,\theta) = \text{acoustic model}.$$

Machine Translation

$Y =$ sentence, $\quad P(Y|\theta) =$ language model,

$X =$ foreign sentence, $\quad P(X|Y,\theta) =$ translation model.

Image Denoising

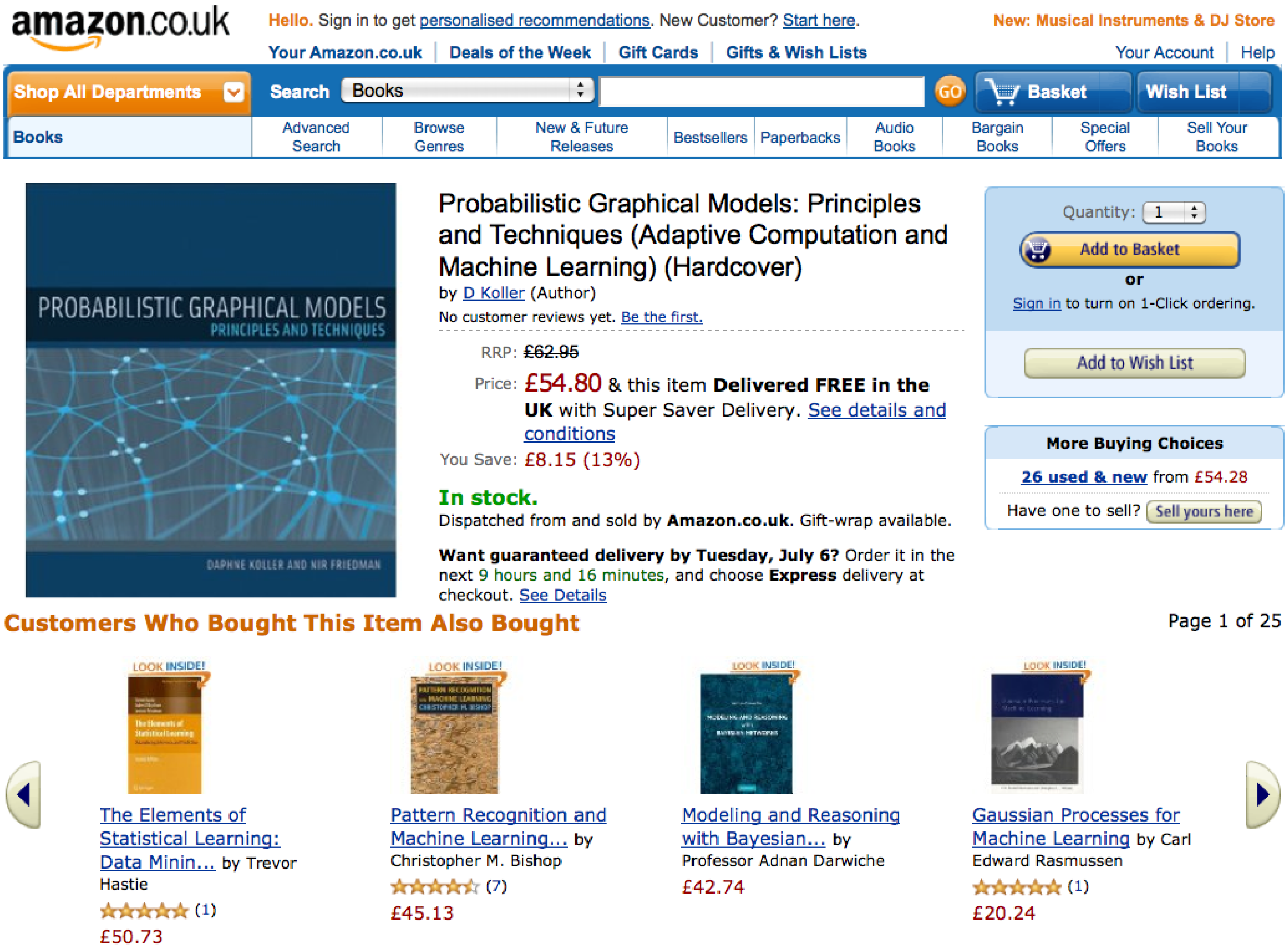
$$
\begin{aligned}
Y &= \text{original image,} \\
X &= \text{noisy image,} \\
P(Y|\theta) &= \text{image model,} \\
P(X|Y,\theta) &= \text{noise model.}
\end{aligned}
$$

Also: deblurring, inpainting, super-resolution...

Simultaneous Localization and Mapping

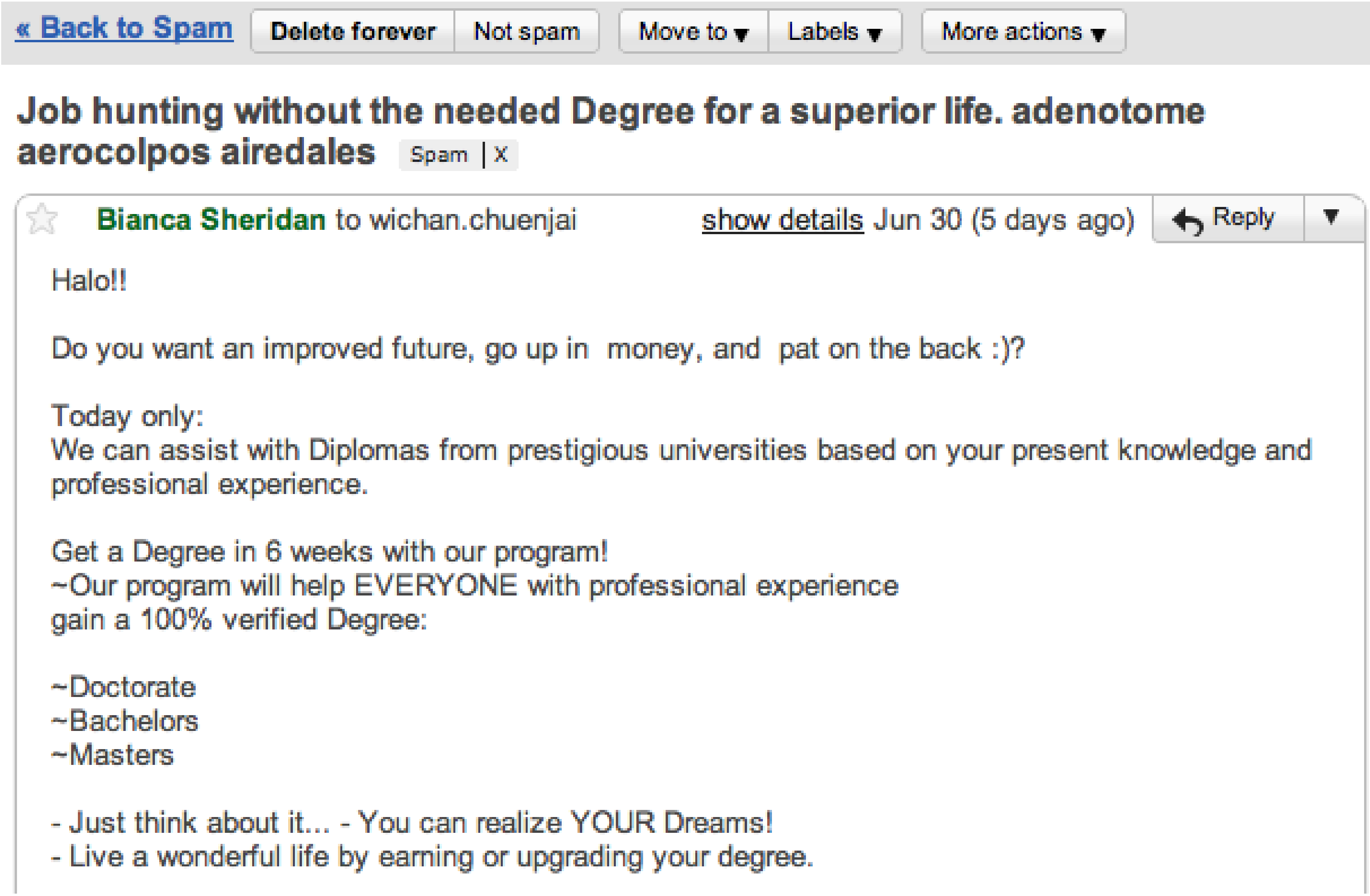
$$Y = \text{map \& location},$$
$$X = \text{sensor readings},$$
$$P(Y|\theta) = \text{map prior},$$
$$P(X|Y,\theta) = \text{observation model}.$$

Collaborative Filtering



$Y =$ user preferences & item features $\qquad X =$ ratings & sales records.

Spam Detection


« Back to Spam | Delete forever | Not spam | Move to ▾ | Labels ▾ | More actions ▾

Job hunting without the needed Degree for a superior life. adenotome aerocolpos airedales Spam | X

Bianca Sheridan to wichan.chuenjai show details Jun 30 (5 days ago) Reply ▾

Halo!!

Do you want an improved future, go up in money, and pat on the back :)?

Today only:
We can assist with Diplomas from prestigious universities based on your present knowledge and professional experience.

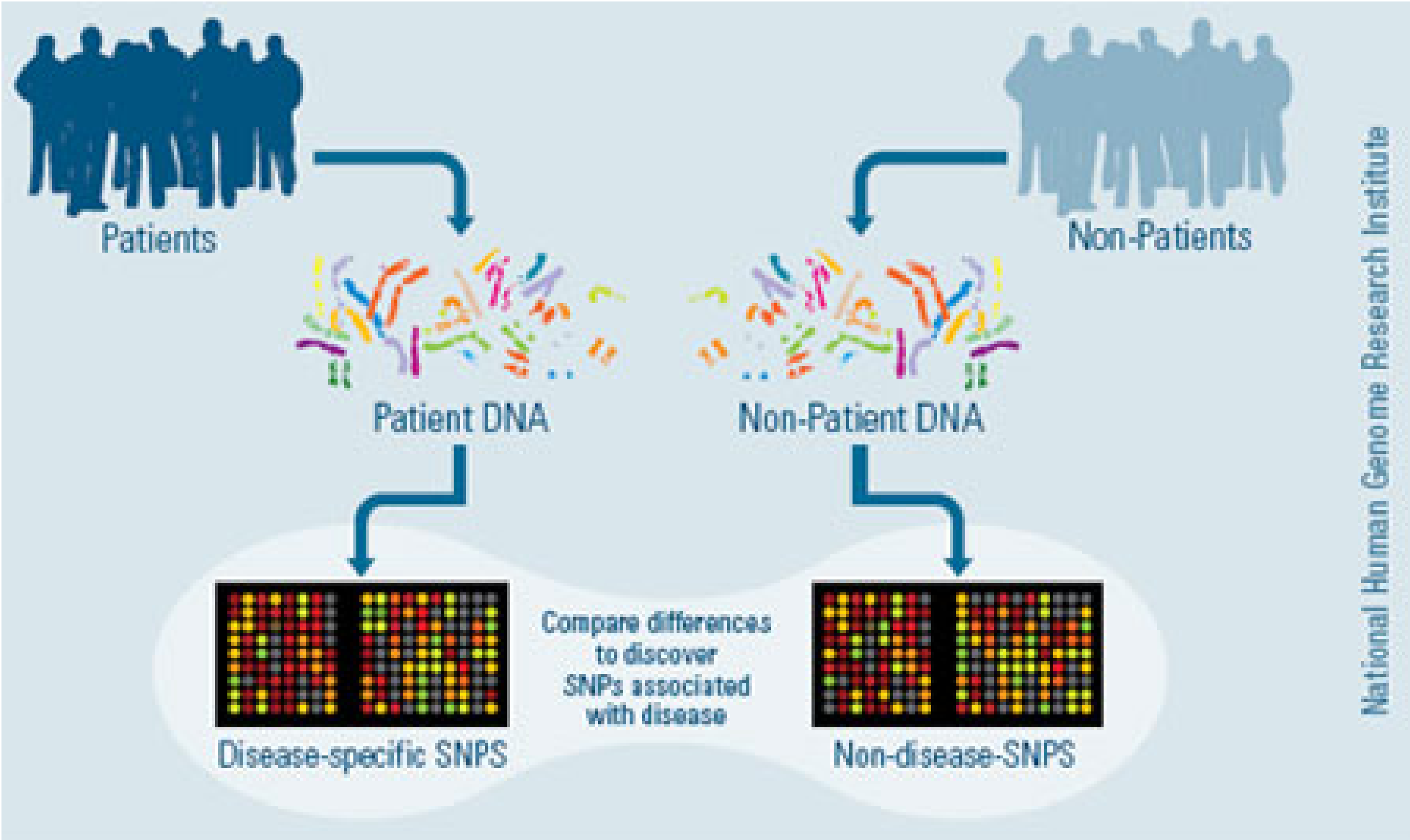
Get a Degree in 6 weeks with our program!
~Our program will help EVERYONE with professional experience
gain a 100% verified Degree:

~Doctorate
~Bachelors
~Masters

- Just think about it... - You can realize YOUR Dreams!
- Live a wonderful life by earning or upgrading your degree.

$Y =$ spam or not? $X =$ text, From:, To: etc.

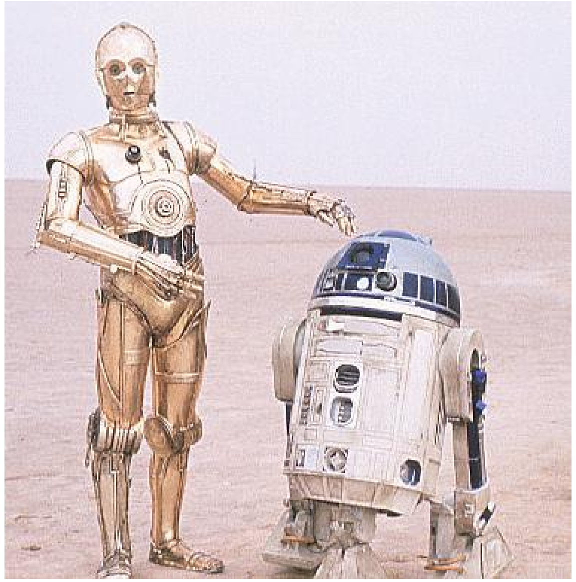
Genome Wide Association Studies



$Y =$ associations, $X =$ DNA, phenotypes, diseases.

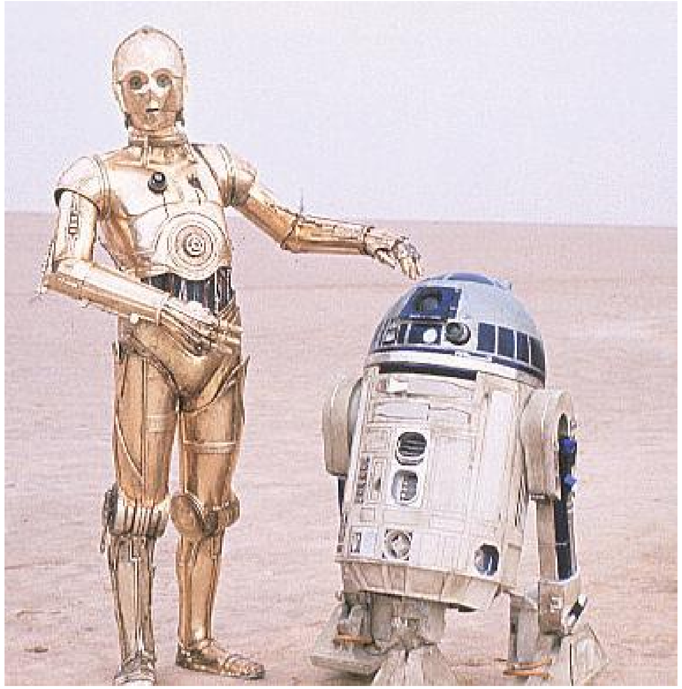
Bayesian theory for representing beliefs

- Goal: to represent the beliefs of learning agents.
- Cox Axioms lead to the following:
  If plausibilities/beliefs are represented by real numbers, then the only reasonable and consistent way to manipulate them is Bayes rule.
- The Dutch Book Theorem:
  If you are willing to bet on your beliefs, then unless they satisfy Bayes rule there will always be a set of bets ("Dutch book") that you would accept which is guaranteed to lose you money, no matter what outcome!
- Frequency vs belief interpretation of probabilities.



Cox Axioms

Consider a robot. In order to behave intelligently the robot should be able to represent beliefs about propositions in the world:



"my charging station is at location (x,y,z)"

"my rangefinder is malfunctioning"

"that stormtrooper is hostile"

We want to represent the strength of these beliefs numerically in the brain of the robot, and we want to know what rules (calculus) we should use to manipulate those beliefs.

Cox Axioms

Let's use $b(x)$ to represent the strength of belief in (plausibility of) proposition x.

$0 \leq b(x) \leq 1$		
$b(x) = 0$	x	is definitely **not true**
$b(x) = 1$	x	is definitely **true**
$b(x\|y)$	strength of belief that x is true given that we know y is true	

Cox Axioms (Desiderata):

- Strengths of belief (degrees of plausibility) are represented by real numbers
- Qualitative correspondence with common sense
- Consistency
  - If a conclusion can be reasoned in more than one way, then every way should lead to the same answer.
  - The robot always takes into account all relevant evidence.
  - Equivalent states of knowledge are represented by equivalent plausibility assignments.

Consequence: Belief functions (e.g. $b(x)$, $b(x|y)$, $b(x,y)$) must satisfy the rules of probability theory, including Bayes rule. (see Jaynes, *Probability Theory: The Logic of Science*)

Dutch Book Theorem

Assume you are willing to accept bets with odds proportional to the strength of your beliefs. That is, $b(x) = 0.9$ implies that you will accept a bet:

$$x \text{ at } 1:9 \Rightarrow \begin{cases} x & \text{is true} \quad \text{win} \quad \geq £1 \\ x & \text{is false} \quad \text{lose} \quad £9 \end{cases}$$

Then, unless your beliefs satisfy the rules of probability theory, including Bayes rule, there exists a set of simultaneous bets (called a Dutch book) which you are willing to accept, and for which **you are guaranteed to lose money, no matter what the outcome**.

Dutch Book Theorem

Assume you are willing to accept bets with odds proportional to the strength of your beliefs. That is, $b(x) = 0.9$ implies that you will accept a bet:

$$x \text{ at } 1:9 \Rightarrow \begin{cases} x & \text{is true} \quad \text{win} \quad \geq £1 \\ x & \text{is false} \quad \text{lose} \quad £9 \end{cases}$$

Then, unless your beliefs satisfy the rules of probability theory, including Bayes rule, there exists a set of simultaneous bets (called a Dutch book) which you are willing to accept, and for which **you are guaranteed to lose money, no matter what the outcome**.

E.g. suppose $A \cap B = \emptyset$, then

Dutch Book Theorem

Assume you are willing to accept bets with odds proportional to the strength of your beliefs. That is, $b(x) = 0.9$ implies that you will accept a bet:

$$x \text{ at } 1:9 \Rightarrow \begin{cases} x & \text{is true} \quad \text{win} \quad \geq £1 \\ x & \text{is false} \quad \text{lose} \quad £9 \end{cases}$$

Then, unless your beliefs satisfy the rules of probability theory, including Bayes rule, there exists a set of simultaneous bets (called a Dutch book) which you are willing to accept, and for which **you are guaranteed to lose money, no matter what the outcome**.

E.g. suppose $A \cap B = \emptyset$, then

$$\left\{ \begin{array}{rl} b(A) & = 0.3 \\ b(B) & = 0.2 \\ b(A \cup B) & = 0.6 \end{array} \right\} \Rightarrow \text{accept the bets} \left\{ \begin{array}{rl} \neg A & \text{at } 3:7 \\ \neg B & \text{at } 2:8 \\ A \cup B & \text{at } 4:6 \end{array} \right\}$$

Dutch Book Theorem

Assume you are willing to accept bets with odds proportional to the strength of your beliefs. That is, $b(x) = 0.9$ implies that you will accept a bet:

$$x \text{ at } 1:9 \Rightarrow \begin{cases} x & \text{is true} \quad \text{win} \quad \geq £1 \\ x & \text{is false} \quad \text{lose} \quad £9 \end{cases}$$

Then, unless your beliefs satisfy the rules of probability theory, including Bayes rule, there exists a set of simultaneous bets (called a Dutch book) which you are willing to accept, and for which **you are guaranteed to lose money, no matter what the outcome**.

E.g. suppose $A \cap B = \emptyset$, then

$$\left\{ \begin{array}{rl} b(A) & = 0.3 \\ b(B) & = 0.2 \\ b(A \cup B) & = 0.6 \end{array} \right\} \Rightarrow \text{accept the bets} \left\{ \begin{array}{rl} \neg A & \text{at } 3:7 \\ \neg B & \text{at } 2:8 \\ A \cup B & \text{at } 4:6 \end{array} \right\}$$

But then:

$$\begin{array}{rcl} \neg A \cap B & \Rightarrow & \text{win } +3-8+4=-1 \\ A \cap \neg B & \Rightarrow & \text{win } -7+2+4=-1 \\ \neg A \cap \neg B & \Rightarrow & \text{win } +3+2-6=-1 \end{array}$$

Dutch Book Theorem

Assume you are willing to accept bets with odds proportional to the strength of your beliefs. That is, $b(x) = 0.9$ implies that you will accept a bet:

$$x \text{ at } 1:9 \Rightarrow \begin{cases} x & \text{is true} \quad \text{win} \quad \geq £1 \\ x & \text{is false} \quad \text{lose} \quad £9 \end{cases}$$

Then, unless your beliefs satisfy the rules of probability theory, including Bayes rule, there exists a set of simultaneous bets (called a Dutch book) which you are willing to accept, and for which **you are guaranteed to lose money, no matter what the outcome**.

E.g. suppose $A \cap B = \emptyset$, then

$$\left\{ \begin{aligned} b(A) &= 0.3 \\ b(B) &= 0.2 \\ b(A \cup B) &= 0.6 \end{aligned} \right\} \Rightarrow \text{accept the bets} \left\{ \begin{aligned} \neg A & \quad \text{at } 3:7 \\ \neg B & \quad \text{at } 2:8 \\ A \cup B & \quad \text{at } 4:6 \end{aligned} \right\}$$

But then:

$$\begin{aligned} \neg A \cap B &\Rightarrow \text{win} \; +3-8+4=-1 \\ A \cap \neg B &\Rightarrow \text{win} \; -7+2+4=-1 \\ \neg A \cap \neg B &\Rightarrow \text{win} \; +3+2-6=-1 \end{aligned}$$

The only way to guard against Dutch books is to ensure that your beliefs are coherent: i.e. satisfy the rules of probability.

Bayesian Learning

Apply the basic rules of probability to learning from data.

Bayesian Learning

Apply the basic rules of probability to learning from data.

- Problem specification:

  Data: $\mathcal{D} = \{x_1, \ldots, x_n\}$ Models: $\mathcal{M}_1$, $\mathcal{M}_2$, etc. Parameters: θ_i (per model)

  Prior probability of models: $P(\mathcal{M}_i)$.

  Prior probabilities of model parameters: $P(\theta_i|\mathcal{M}_i)$

  Model of data given parameters (likelihood model): $P(X, Y|\theta_i, \mathcal{M}_i)$

Bayesian Learning

Apply the basic rules of probability to learning from data.

- Problem specification:

  Data: $\mathcal{D} = \{x_1, \ldots, x_n\}$ Models: $\mathcal{M}_1$, $\mathcal{M}_2$, etc. Parameters: θ_i (per model)

  Prior probability of models: $P(\mathcal{M}_i)$.

  Prior probabilities of model parameters: $P(\theta_i|\mathcal{M}_i)$

  Model of data given parameters (likelihood model): $P(X, Y|\theta_i, \mathcal{M}_i)$

- Data probability (likelihood)

$$P(\mathcal{D}|\theta_i, \mathcal{M}_i) = \prod_{i=1}^{n} \sum_{y_i} P(x_i, y_i|\theta_i, \mathcal{M}_i) \equiv \mathcal{L}_i(\theta_i)$$

  provided the data are independently and identically distributed (iid).

Bayesian Learning

Apply the basic rules of probability to learning from data.

- Problem specification:

  Data: $\mathcal{D} = \{x_1, \ldots, x_n\}$ Models: $\mathcal{M}_1$, $\mathcal{M}_2$, etc. Parameters: θ_i (per model)

  Prior probability of models: $P(\mathcal{M}_i)$.

  Prior probabilities of model parameters: $P(\theta_i|\mathcal{M}_i)$

  Model of data given parameters (likelihood model): $P(X, Y|\theta_i, \mathcal{M}_i)$

- Data probability (likelihood)

$$P(\mathcal{D}|\theta_i, \mathcal{M}_i) = \prod_{i=1}^{n} \sum_{y_i} P(x_i, y_i|\theta_i, \mathcal{M}_i) \equiv \mathcal{L}_i(\theta_i)$$

  provided the data are independently and identically distributed (iid).

- Parameter learning (posterior):

$$P(\theta_i|\mathcal{D}, \mathcal{M}_i) = \frac{P(\mathcal{D}|\theta_i, \mathcal{M}_i)P(\theta_i|\mathcal{M}_i)}{P(\mathcal{D}|\mathcal{M}_i)}; \quad P(\mathcal{D}|\mathcal{M}_i) = \int d\theta_i \; P(\mathcal{D}|\theta_i, \mathcal{M}_i)P(\theta|\mathcal{M}_i)$$

Bayesian Learning

Apply the basic rules of probability to learning from data.

- Problem specification:

  Data: $\mathcal{D} = \{x_1, \ldots, x_n\}$ Models: $\mathcal{M}_1$, $\mathcal{M}_2$, etc. Parameters: θ_i (per model)

  Prior probability of models: $P(\mathcal{M}_i)$.

  Prior probabilities of model parameters: $P(\theta_i|\mathcal{M}_i)$

  Model of data given parameters (likelihood model): $P(X, Y|\theta_i, \mathcal{M}_i)$

- Data probability (likelihood)

$$P(\mathcal{D}|\theta_i, \mathcal{M}_i) = \prod_{i=1}^{n} \sum_{y_i} P(x_i, y_i|\theta_i, \mathcal{M}_i) \equiv \mathcal{L}_i(\theta_i)$$

  provided the data are independently and identically distributed (iid).

- Parameter learning (posterior):

$$P(\theta_i|\mathcal{D}, \mathcal{M}_i) = \frac{P(\mathcal{D}|\theta_i, \mathcal{M}_i)P(\theta_i|\mathcal{M}_i)}{P(\mathcal{D}|\mathcal{M}_i)}; \quad P(\mathcal{D}|\mathcal{M}_i) = \int d\theta_i \; P(\mathcal{D}|\theta_i, \mathcal{M}_i)P(\theta|\mathcal{M}_i)$$

  $P(\mathcal{D}|\mathcal{M}_i)$ is called the marginal likelihood or evidence for $\mathcal{M}_i$. It is proportional to the posterior probability model $\mathcal{M}_i$ being the one that generated the data.

- Model selection:

$$P(\mathcal{M}_i|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M}_i)P(\mathcal{M}_i)}{P(\mathcal{D})}$$

Bayesian Learning: A coin toss example

Coin toss: One parameter q — the odds of obtaining *heads*

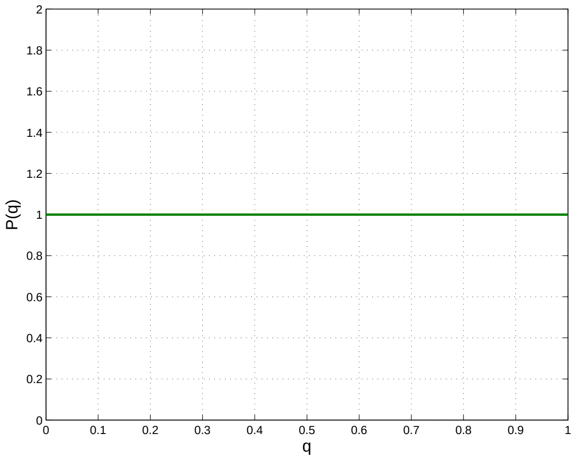
So our space of models is the set of distributions over $q \in [0, 1]$.

Bayesian Learning: A coin toss example

Coin toss: One parameter q — the odds of obtaining *heads*

So our space of models is the set of distributions over $q \in [0, 1]$.

Learner A believes model $\mathcal{M}_\mathsf{A}$: all values of q are equally plausible;



A

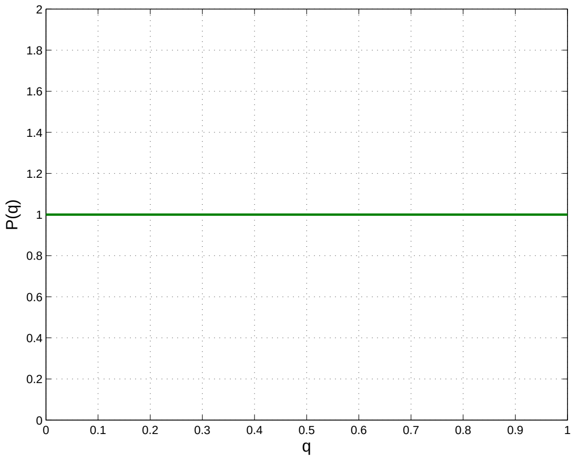
Bayesian Learning: A coin toss example

Coin toss: One parameter q — the odds of obtaining *heads*

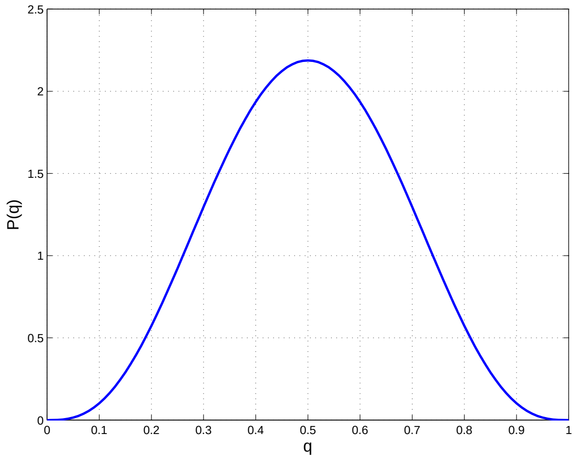
So our space of models is the set of distributions over $q \in [0, 1]$.

Learner A believes model $\mathcal{M}_{\text{A}}$: all values of q are equally plausible;

Learner B believes model $\mathcal{M}_{\text{B}}$: more plausible that the coin is “fair” ($q \approx 0.5$) than “biased”.



A



B

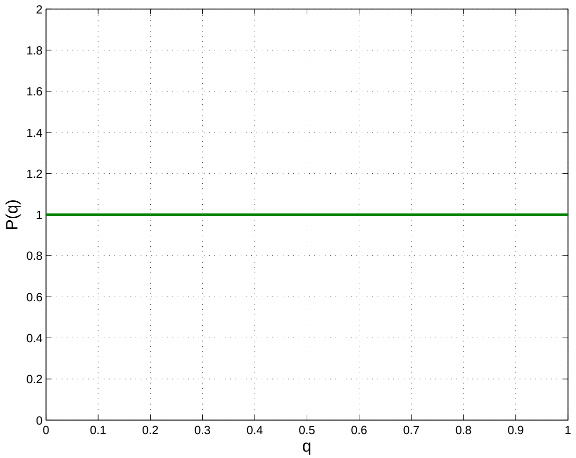
Bayesian Learning: A coin toss example

Coin toss: One parameter q — the odds of obtaining *heads*

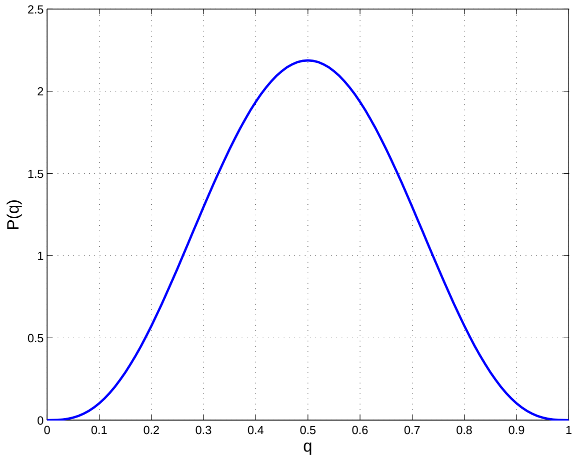
So our space of models is the set of distributions over $q \in [0, 1]$.

Learner A believes model $\mathcal{M}_\text{A}$: all values of q are equally plausible;

Learner B believes model $\mathcal{M}_\text{B}$: more plausible that the coin is "fair" ($q \approx 0.5$) than "biased".



A: $\alpha_1 = \alpha_2 = 1.0$



B: $\alpha_1 = \alpha_2 = 4.0$

Both prior beliefs can be described by the Beta distribution:

$$p(q|\alpha_1, \alpha_2) = \frac{q^{(\alpha_1 - 1)}(1-q)^{(\alpha_2 - 1)}}{B(\alpha_1, \alpha_2)} = \mathrm{Beta}(q|\alpha_1, \alpha_2)$$

where B is the (beta) function which normalizes the distribution:

$$B(\alpha_1, \alpha_2) = \int_0^1 t^{\alpha_1 - 1}(1-t)^{\alpha_2 - 1}\, dt = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}$$

Bayesian Learning: A coin toss example

Now we observe a toss. Two possible outcomes:

$$p(\mathsf{H}|q) = q \qquad p(\mathsf{T}|q) = 1 - q$$

Bayesian Learning: A coin toss example

Now we observe a toss. Two possible outcomes:

$$p(\mathsf{H}|q) = q \qquad p(\mathsf{T}|q) = 1 - q$$

Suppose our single coin toss comes out heads.

Bayesian Learning: A coin toss example

Now we observe a toss. Two possible outcomes:

$$p(\mathsf{H}|q) = q \qquad p(\mathsf{T}|q) = 1 - q$$

Suppose our single coin toss comes out heads.

The probability of the observed data (likelihood) is:

$$p(\mathsf{H}|q) = q$$

Bayesian Learning: A coin toss example

Now we observe a toss. Two possible outcomes:

$$p(\mathsf{H}|q) = q \qquad p(\mathsf{T}|q) = 1 - q$$

Suppose our single coin toss comes out heads.

The probability of the observed data (likelihood) is:

$$p(\mathsf{H}|q) = q$$

Using Bayes Rule, we multiply the prior, $p(q)$ by the likelihood and renormalise to get the posterior probability:

$$p(q|\mathsf{H}) = \frac{p(q)p(\mathsf{H}|q)}{p(\mathsf{H})}$$

Bayesian Learning: A coin toss example

Now we observe a toss. Two possible outcomes:

$$p(\mathsf{H}|q) = q \qquad p(\mathsf{T}|q) = 1 - q$$

Suppose our single coin toss comes out heads.

The probability of the observed data (likelihood) is:

$$p(\mathsf{H}|q) = q$$

Using Bayes Rule, we multiply the prior, $p(q)$ by the likelihood and renormalise to get the posterior probability:

$$p(q|\mathsf{H}) \; = \; \frac{p(q)p(\mathsf{H}|q)}{p(\mathsf{H})} \propto q \; \mathrm{Beta}(q|\alpha_1, \alpha_2)$$

Bayesian Learning: A coin toss example

Now we observe a toss. Two possible outcomes:

$$p(\mathsf{H}|q) = q \qquad p(\mathsf{T}|q) = 1 - q$$

Suppose our single coin toss comes out heads.

The probability of the observed data (likelihood) is:

$$p(\mathsf{H}|q) = q$$

Using Bayes Rule, we multiply the prior, $p(q)$ by the likelihood and renormalise to get the posterior probability:

$$p(q|\mathsf{H}) \;=\; \frac{p(q)p(\mathsf{H}|q)}{p(\mathsf{H})} \propto q \; \mathrm{Beta}(q|\alpha_1, \alpha_2)$$

$$\propto \; q \; q^{(\alpha_1 - 1)}(1-q)^{(\alpha_2 - 1)}$$

Bayesian Learning: A coin toss example

Now we observe a toss. Two possible outcomes:

$$p(\mathsf{H}|q) = q \qquad p(\mathsf{T}|q) = 1 - q$$

Suppose our single coin toss comes out heads.

The probability of the observed data (likelihood) is:

$$p(\mathsf{H}|q) = q$$

Using Bayes Rule, we multiply the prior, $p(q)$ by the likelihood and renormalise to get the posterior probability:

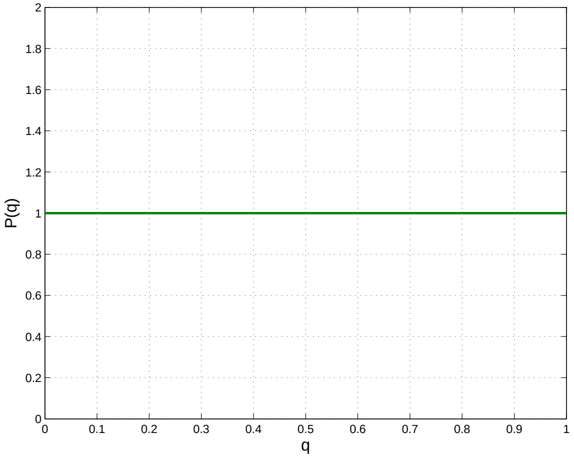
$$p(q|\mathsf{H}) = \frac{p(q)p(\mathsf{H}|q)}{p(\mathsf{H})} \propto q \ \mathrm{Beta}(q|\alpha_1, \alpha_2)$$

$$\propto q \ q^{(\alpha_1 - 1)}(1-q)^{(\alpha_2 - 1)} = \mathrm{Beta}(q|\alpha_1 + 1, \alpha_2)$$

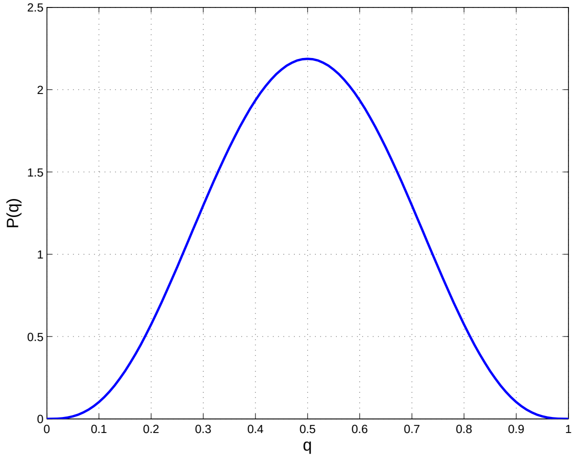
Bayesian Learning: A coin toss example

A | B

Prior

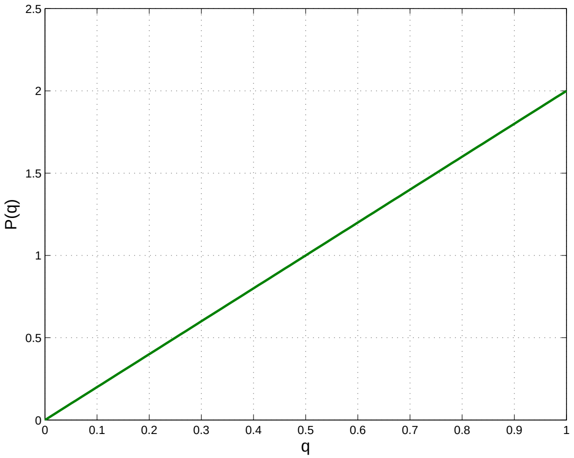


$\text{Beta}(q|1,1)$

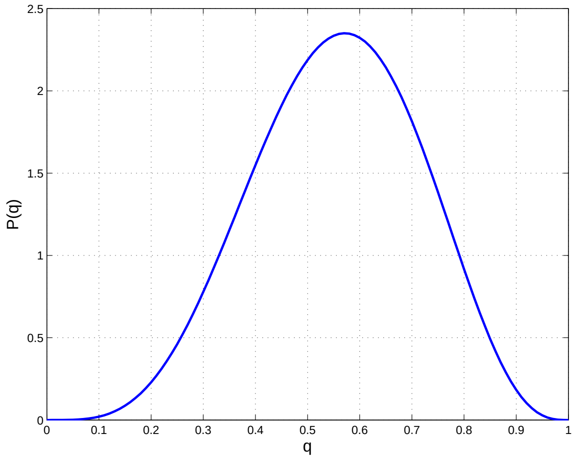


$\text{Beta}(q|4,4)$

Posterior



$\text{Beta}(q|2,1)$



$\text{Beta}(q|5,4)$

Bayesian Learning: A coin toss example

What about multiple tosses?

Bayesian Learning: A coin toss example

What about multiple tosses? Suppose we observe $\mathcal{D} = \{$ H H T H T T $\}$:

$$p(\{\text{ H H T H T T }\}|q) = qq(1-q)q(1-q)(1-q) = q^3(1-q)^3$$

Bayesian Learning: A coin toss example

What about multiple tosses? Suppose we observe $\mathcal{D} = \{$ H H T H T T $\}$:

$$p(\{ \text{H H T H T T} \}|q) = qq(1-q)q(1-q)(1-q) = q^3(1-q)^3$$

This is still straightforward:

Bayesian Learning: A coin toss example

What about multiple tosses? Suppose we observe $\mathcal{D} = \{$ H H T H T T $\}$:

$$p(\{\text{ H H T H T T }\}|q) = qq(1-q)q(1-q)(1-q) = q^3(1-q)^3$$

This is still straightforward:

$$p(q|\mathcal{D}) \;=\; \frac{p(q)p(\mathcal{D}|q)}{p(\mathcal{D})}$$

Bayesian Learning: A coin toss example

What about multiple tosses? Suppose we observe $\mathcal{D} = \{$ H H T H T T $\}$:

$$p(\{\text{ H H T H T T }\}|q) = qq(1-q)q(1-q)(1-q) = q^3(1-q)^3$$

This is still straightforward:

$$p(q|\mathcal{D}) \; = \; \frac{p(q)p(\mathcal{D}|q)}{p(\mathcal{D})} \propto q^3(1-q)^3 \, \mathrm{Beta}(q|\alpha_1, \alpha_2)$$

Bayesian Learning: A coin toss example

What about multiple tosses? Suppose we observe $\mathcal{D} = \{ \text{H H T H T T} \}$:

$$p(\{ \text{H H T H T T} \}|q) = qq(1-q)q(1-q)(1-q) = q^3(1-q)^3$$

This is still straightforward:

$$p(q|\mathcal{D}) = \frac{p(q)p(\mathcal{D}|q)}{p(\mathcal{D})} \propto q^3(1-q)^3 \, \text{Beta}(q|\alpha_1, \alpha_2)$$

$$\propto \text{Beta}(q|\alpha_1 + 3, \alpha_2 + 3)$$

Bayesian Learning: A coin toss example

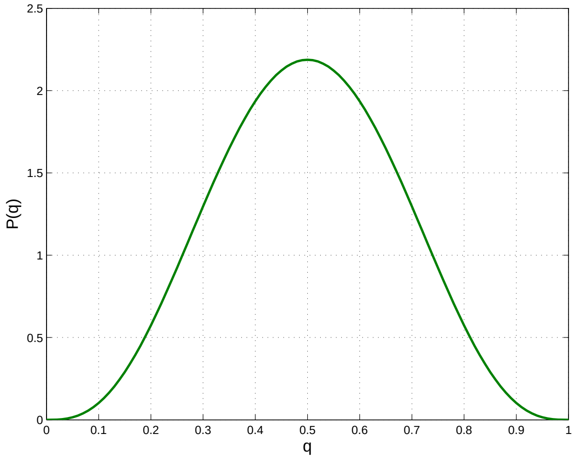
What about multiple tosses? Suppose we observe $\mathcal{D} = \{\text{ H H T H T T }\}$:

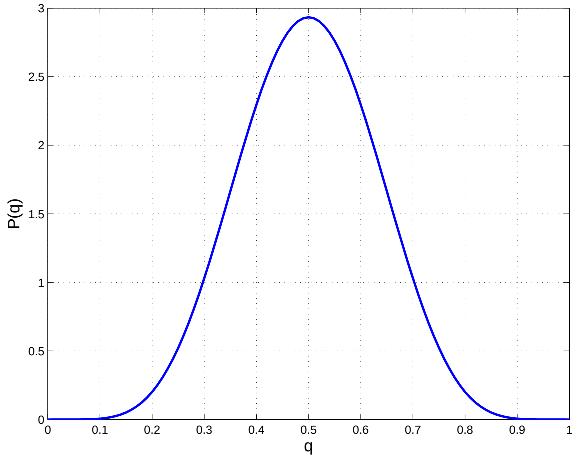
$$p(\{\text{ H H T H T T }\}|q) = qq(1-q)q(1-q)(1-q) = q^3(1-q)^3$$

This is still straightforward:

$$p(q|\mathcal{D}) = \frac{p(q)p(\mathcal{D}|q)}{p(\mathcal{D})} \propto q^3(1-q)^3 \, \mathrm{Beta}(q|\alpha_1, \alpha_2)$$

$$\propto \mathrm{Beta}(q|\alpha_1 + 3, \alpha_2 + 3)$$





Conjugate Priors

Updating the prior to form the posterior was particularly easy in these examples.

Conjugate Priors

Updating the prior to form the posterior was particularly easy in these examples. This is because we used a conjugate prior for an exponential family likelihood.

Conjugate Priors

Updating the prior to form the posterior was particularly easy in these examples. This is because we used a conjugate prior for an exponential family likelihood.

Exponential family distributions take the form:

$$P(x|\theta) = g(\theta) f(x) e^{\boldsymbol{\phi}(\theta)^{\mathsf{T}} \mathbf{T}(x)}$$

with $g(\theta)$ the normalising constant.

Conjugate Priors

Updating the prior to form the posterior was particularly easy in these examples. This is because we used a conjugate prior for an exponential family likelihood.

Exponential family distributions take the form:

$$P(x|\theta) = g(\theta) f(x) e^{\boldsymbol{\phi}(\theta)^{\mathsf{T}} \mathbf{T}(x)}$$

with $g(\theta)$ the normalising constant. Given n iid observations,

$$P(\{x_i\}|\theta) = \prod_i P(x_i|\theta) = g(\theta)^n e^{\boldsymbol{\phi}(\theta)^{\mathsf{T}} \left(\sum_i \mathbf{T}(x_i) \right)} \prod_i f(x_i)$$

Conjugate Priors

Updating the prior to form the posterior was particularly easy in these examples. This is because we used a conjugate prior for an exponential family likelihood.

Exponential family distributions take the form:

$$P(x|\theta) = g(\theta) f(x) e^{\boldsymbol{\phi}(\theta)^\mathsf{T} \mathbf{T}(x)}$$

with $g(\theta)$ the normalising constant. Given n iid observations,

$$P(\{x_i\}|\theta) = \prod_i P(x_i|\theta) = g(\theta)^n e^{\boldsymbol{\phi}(\theta)^\mathsf{T} \left(\sum_i \mathbf{T}(x_i) \right)} \prod_i f(x_i)$$

Thus, if the prior takes the conjugate form

$$P(\theta) = F(\boldsymbol{\tau}, \nu) g(\theta)^\nu e^{\boldsymbol{\phi}(\theta)^\mathsf{T} \boldsymbol{\tau}}$$

with $F(\boldsymbol{\tau}, \nu)$ the normaliser

Conjugate Priors

Updating the prior to form the posterior was particularly easy in these examples. This is because we used a conjugate prior for an exponential family likelihood.

Exponential family distributions take the form:

$$P(x|\theta) = g(\theta) f(x) e^{\boldsymbol{\phi}(\theta)^{\mathsf{T}} \mathbf{T}(x)}$$

with $g(\theta)$ the normalising constant. Given n iid observations,

$$P(\{x_i\}|\theta) = \prod_i P(x_i|\theta) = g(\theta)^n e^{\boldsymbol{\phi}(\theta)^{\mathsf{T}} \left(\sum_i \mathbf{T}(x_i)\right)} \prod_i f(x_i)$$

Thus, if the prior takes the conjugate form

$$P(\theta) = F(\boldsymbol{\tau}, \nu) g(\theta)^{\nu} e^{\boldsymbol{\phi}(\theta)^{\mathsf{T}} \boldsymbol{\tau}}$$

with $F(\boldsymbol{\tau}, \nu)$ the normaliser, then the posterior is

$$P(\theta|\{x_i\}) \propto P(\{x_i\}|\theta) P(\theta) \propto g(\theta)^{\nu+n} e^{\boldsymbol{\phi}(\theta)^{\mathsf{T}} \left(\boldsymbol{\tau} + \sum_i \mathbf{T}(x_i)\right)}$$

with the normaliser given by $F\left(\boldsymbol{\tau} + \sum_i \mathbf{T}(x_i), \nu + n\right)$.

Conjugate Priors

The posterior given an exponential family likelihood and conjugate prior is:

$$P(\theta|\{x_i\}) = F\Big(\boldsymbol{\tau} + \sum_i \mathbf{T}(x_i), \nu + n\Big) g(\theta)^{\nu+n} \exp\left[\boldsymbol{\phi}(\theta)^{\mathsf{T}}\Big(\boldsymbol{\tau} + \sum_i \mathbf{T}(x_i)\Big)\right]$$

Here,

$\boldsymbol{\phi}(\theta)$	is the vector of natural parameters
$\sum_i \mathbf{T}(x_i)$	is the vector of sufficient statistics
$\boldsymbol{\tau}$	are pseudo-observations which define the prior
ν	is the scale of the prior (need not be an integer)

As new data come in, each one increments the sufficient statistics vector and the scale to define the posterior.

Conjugacy in the coin toss example

Distributions are not always written in their natural exponential form.

Conjugacy in the coin toss example

Distributions are not always written in their natural exponential form.

The Bernoulli distribution (a single coin flip) with parameter q and observation $x \in \{0, 1\}$, can be written:

$$
\begin{aligned}
P(x|q) &= q^x (1-q)^{(1-x)} \\
&= e^{x \log q + (1-x) \log(1-q)} \\
&= e^{\log(1-q) + x \log(q/(1-q))} \\
&= (1-q) e^{\log(q/(1-q)) x}
\end{aligned}
$$

Conjugacy in the coin toss example

Distributions are not always written in their natural exponential form.
The Bernoulli distribution (a single coin flip) with parameter q and observation $x \in \{0, 1\}$, can be written:

$$\begin{aligned} P(x|q) &= q^x (1-q)^{(1-x)} \\ &= e^{x \log q + (1-x) \log(1-q)} \\ &= e^{\log(1-q) + x \log(q/(1-q))} \\ &= (1-q) e^{\log(q/(1-q)) x} \end{aligned}$$

So the natural parameter is the log odds $\log(q/(1-q))$, and the sufficient statisticss (for multiple tosses) is the number of heads.

Conjugacy in the coin toss example

Distributions are not always written in their natural exponential form.
The Bernoulli distribution (a single coin flip) with parameter q and observation $x \in \{0,1\}$, can be written:

$$\begin{aligned} P(x|q) &= q^x(1-q)^{(1-x)} \\ &= e^{x\log q + (1-x)\log(1-q)} \\ &= e^{\log(1-q) + x\log(q/(1-q))} \\ &= (1-q)e^{\log(q/(1-q))x} \end{aligned}$$

So the natural parameter is the log odds $\log(q/(1-q))$, and the sufficient statisticss (for multiple tosses) is the number of heads.
The conjugate prior is

$$\begin{aligned} P(q) &= F(\tau,\nu)\ (1-q)^{\nu} e^{\log(q/(1-q))\tau} \\ &= F(\tau,\nu)\ (1-q)^{\nu} e^{\tau\log q - \tau\log(1-q)} \\ &= F(\tau,\nu)\ (1-q)^{\nu-\tau} q^{\tau} \end{aligned}$$

which has the form of the Beta distribution $\Rightarrow F(\tau,\nu) = 1/B(\tau+1, \nu-\tau+1)$.

Conjugacy in the coin toss example

Distributions are not always written in their natural exponential form.

The Bernoulli distribution (a single coin flip) with parameter q and observation $x \in \{0, 1\}$, can be written:

$$\begin{aligned} P(x|q) &= q^x (1-q)^{(1-x)} \\ &= e^{x \log q + (1-x)\log(1-q)} \\ &= e^{\log(1-q) + x \log(q/(1-q))} \\ &= (1-q) e^{\log(q/(1-q))x} \end{aligned}$$

So the natural parameter is the log odds $\log(q/(1-q))$, and the sufficient statisticss (for multiple tosses) is the number of heads.

The conjugate prior is

$$\begin{aligned} P(q) &= F(\tau, \nu)\ (1-q)^\nu e^{\log(q/(1-q))\tau} \\ &= F(\tau, \nu)\ (1-q)^\nu e^{\tau \log q - \tau \log(1-q)} \\ &= F(\tau, \nu)\ (1-q)^{\nu - \tau} q^\tau \end{aligned}$$

which has the form of the Beta distribution $\Rightarrow F(\tau, \nu) = 1/B(\tau + 1, \nu - \tau + 1)$.

In general, then, the posterior will be $P(q|\{x_i\}) \propto q^{\alpha_1 - 1}(1-q)^{\alpha_2 - 1} = \mathrm{Beta}(q|\alpha_1, \alpha_2)$, with

$$\alpha_1 = 1 + \tau + \textstyle\sum_i x_i \qquad \alpha_2 = 1 + (\nu + n) - \left(\tau + \textstyle\sum_i x_i\right)$$

Conjugacy in the coin toss example

Distributions are not always written in their natural exponential form.
The Bernoulli distribution (a single coin flip) with parameter q and observation $x \in \{0, 1\}$, can be written:

$$\begin{aligned} P(x|q) &= q^x (1-q)^{(1-x)} \\ &= e^{x \log q + (1-x)\log(1-q)} \\ &= e^{\log(1-q) + x \log(q/(1-q))} \\ &= (1-q) e^{\log(q/(1-q))x} \end{aligned}$$

So the natural parameter is the log odds $\log(q/(1-q))$, and the sufficient statisticss (for multiple tosses) is the number of heads.
The conjugate prior is

$$\begin{aligned} P(q) &= F(\tau, \nu)\ (1-q)^{\nu} e^{\log(q/(1-q))\tau} \\ &= F(\tau, \nu)\ (1-q)^{\nu} e^{\tau \log q - \tau \log(1-q)} \\ &= F(\tau, \nu)\ (1-q)^{\nu-\tau} q^{\tau} \end{aligned}$$

which has the form of the Beta distribution $\Rightarrow F(\tau, \nu) = 1/B(\tau + 1, \nu - \tau + 1)$.
In general, then, the posterior will be $P(q|\{x_i\}) \propto q^{\alpha_1 - 1}(1-q)^{\alpha_2 - 1} = \mathrm{Beta}(q|\alpha_1, \alpha_2)$, with

$$\alpha_1 = 1 + \tau + \textstyle\sum_i x_i \qquad\qquad \alpha_2 = 1 + (\nu + n) - \left(\tau + \textstyle\sum_i x_i\right)$$

If we observe a head, we add 1 to the sufficient statistic $\sum x_i$, and also 1 to the count n.
This increments α_1. If we observe a tail we add 1 to n, but not to $\sum x_i$, incrementing α_2.

Model selection in coin toss example

We have seen how to update posteriors within each model. To study the choice of model, consider two more extreme models: “fair” and “bent”.

Model selection in coin toss example

We have seen how to update posteriors within each model. To study the choice of model, consider two more extreme models: “fair” and “bent”. A priori, we may think that “fair” is more probable, eg:

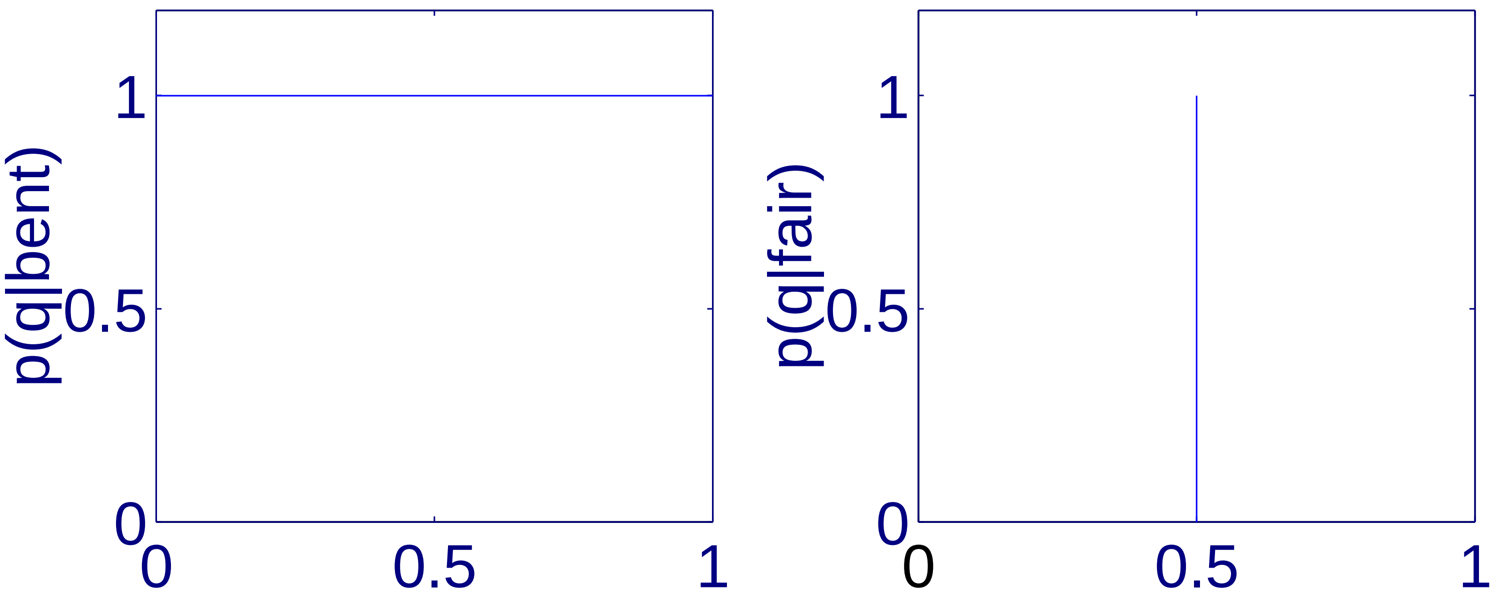
$$p(\text{fair}) = 0.8, \qquad p(\text{bent}) = 0.2$$

Model selection in coin toss example

We have seen how to update posteriors within each model. To study the choice of model, consider two more extreme models: “fair” and “bent”. A priori, we may think that “fair” is more probable, eg:

$$p(\text{fair}) = 0.8, \qquad p(\text{bent}) = 0.2$$

For the bent coin, we assume all parameter values are equally likely, whilst the fair coin has a fixed probability:

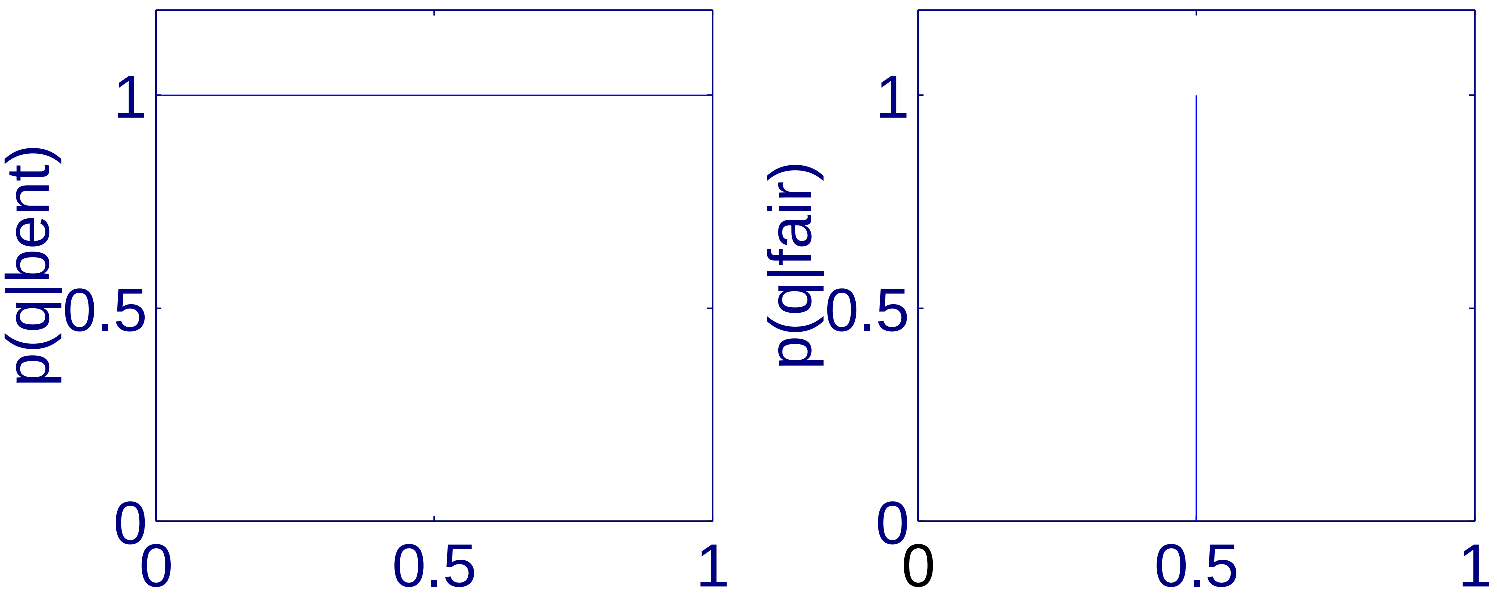


Model selection in coin toss example

We have seen how to update posteriors within each model. To study the choice of model, consider two more extreme models: "fair" and "bent". A priori, we may think that "fair" is more probable, eg:

$$p(\text{fair}) = 0.8, \qquad p(\text{bent}) = 0.2$$

For the bent coin, we assume all parameter values are equally likely, whilst the fair coin has a fixed probability:



We make 10 tosses, and get: $\mathcal{D} = (\text{T H T H T T T T T T})$.

Model selection in coin toss example

Which model should we prefer *a posteriori* (i.e. after seeing the data)?

Model selection in coin toss example

Which model should we prefer *a posteriori* (i.e. after seeing the data)?

The evidence for the fair model is:

$$P(\mathcal{D}|\text{fair}) = (1/2)^{10} \approx 0.001$$

Model selection in coin toss example

Which model should we prefer *a posteriori* (i.e. after seeing the data)?

The evidence for the fair model is:

$$P(\mathcal{D}|\text{fair}) = (1/2)^{10} \approx 0.001$$

and for the bent model is:

$$P(\mathcal{D}|\text{bent}) = \int dq\ P(\mathcal{D}|q, \text{bent})p(q|\text{bent}) = \int dq\ q^2(1-q)^8 = \mathrm{B}(3,9) \approx 0.002$$

Model selection in coin toss example

Which model should we prefer *a posteriori* (i.e. after seeing the data)?

The evidence for the fair model is:

$$P(\mathcal{D}|\text{fair}) = (1/2)^{10} \approx 0.001$$

and for the bent model is:

$$P(\mathcal{D}|\text{bent}) = \int dq\ P(\mathcal{D}|q, \text{bent})p(q|\text{bent}) = \int dq\ q^2(1-q)^8 = \mathrm{B}(3,9) \approx 0.002$$

Thus, the posterior for the models, by Bayes rule:

$$P(\text{fair}|\mathcal{D}) \propto 0.0008, \qquad P(\text{bent}|\mathcal{D}) \propto 0.0004,$$

ie, a two-thirds probability that the coin is fair.

Model selection in coin toss example

Which model should we prefer *a posteriori* (i.e. after seeing the data)?

The evidence for the fair model is:

$$P(\mathcal{D}|\text{fair}) = (1/2)^{10} \approx 0.001$$

and for the bent model is:

$$P(\mathcal{D}|\text{bent}) = \int dq\ P(\mathcal{D}|q, \text{bent}) p(q|\text{bent}) = \int dq\ q^2(1-q)^8 = \mathrm{B}(3,9) \approx 0.002$$

Thus, the posterior for the models, by Bayes rule:

$$P(\text{fair}|\mathcal{D}) \propto 0.0008, \qquad P(\text{bent}|\mathcal{D}) \propto 0.0004,$$

ie, a two-thirds probability that the coin is fair.

How do we make predictions?

Model selection in coin toss example

Which model should we prefer *a posteriori* (i.e. after seeing the data)?

The evidence for the fair model is:

$$P(\mathcal{D}|\text{fair}) = (1/2)^{10} \approx 0.001$$

and for the bent model is:

$$P(\mathcal{D}|\text{bent}) = \int dq\ P(\mathcal{D}|q, \text{bent})p(q|\text{bent}) = \int dq\ q^2(1-q)^8 = \mathrm{B}(3,9) \approx 0.002$$

Thus, the posterior for the models, by Bayes rule:

$$P(\text{fair}|\mathcal{D}) \propto 0.0008, \qquad P(\text{bent}|\mathcal{D}) \propto 0.0004,$$

ie, a two-thirds probability that the coin is fair.

How do we make predictions? Could choose the fair model (model selection).
Or could weight the predictions from each model by their probability (model averaging).
Probability of H at next toss is:

$$P(\text{H}|\mathcal{D}) = P(\text{H}|\text{fair})P(\text{fair}|\mathcal{D}) + P(\text{H}|\text{bent})P(\text{bent}|\mathcal{D}) = \frac{2}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{3}{12} = \frac{5}{12}.$$

Bayesian Learning

If an agent is learning parameters, it could report different aspects of the posterior (or likelihood).

- Bayesian Learning: Assumes a prior over the model parameters. Computes the posterior distribution of the parameters: $P(\theta|\mathcal{D})$.
- Maximum a Posteriori (MAP) Learning: Assumes a prior over the model parameters $P(\theta)$. Finds a parameter setting that maximises the posterior: $P(\theta|\mathcal{D}) \propto P(\theta)P(\mathcal{D}|\theta)$.
- Maximum Likelihood (ML) Learning: Does not assume a prior over the model parameters. Finds a parameter setting that maximises the likelihood function: $P(\mathcal{D}|\theta)$.

Choosing between these and other alternatives may be a matter of definition, of goals, or of practicality

In practice (outside the exponential family), the Bayesian ideal may be computationally challenging, and may need to be approximated at best.

We will return to the Bayesian formulation on Thursday and Friday. For Tuesday and Wednesday we will look at ML and MAP learning in more complex graphical models.

End Notes

The following notes by Sam Roweis are quite useful:

Matrix identities and matrix derivatives:
`http://www.cs.toronto.edu/~roweis/notes/matrixid.pdf`

Gaussian identities:
`http://www.cs.toronto.edu/~roweis/notes/gaussid.pdf`

List of some useful exponential family distributions:
`http://www.cse.buffalo.edu/faculty/mbeal/papers/vbqref.pdf`

Here is a useful statistics / pattern recognition glossary:
`http://research.microsoft.com/~minka/statlearn/glossary/`

Tom Minka's in-depth notes on matrix algebra:
`http://research.microsoft.com/~minka/papers/matrix/`

End Notes