SC4/SM8 Advanced Topics in Statistical Machine Learning

Bayesian Learning

Dino Sejdinovic
Department of Statistics
Oxford

Slides and other materials available at:
http://www.stats.ox.ac.uk/~sejdinov/atsml/
Bayesian Learning

Review of Bayesian Inference

The Bayesian Learning Framework

- Bayesian learning: **treat parameter vector** $\theta$ **as a random variable**: process of learning is then **computation of the posterior distribution** $p(\theta|D)$.
- In addition to the likelihood $p(D|\theta)$ need to specify a **prior distribution** $p(\theta)$.
- Posterior distribution is then given by the **Bayes Theorem**:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- **Likelihood**: $p(D|\theta)$
- **Prior**: $p(\theta)$
- **Posterior**: $p(\theta|D)$
- **Marginal likelihood**: $p(D) = \int_\Theta p(D|\theta)p(\theta)d\theta$

- Summarizing the posterior:
  - **Posterior mode**: $\hat{\theta}^{\text{MAP}} = \arg\max_{\theta \in \Theta} p(\theta|D)$ (maximum a posteriori).
  - **Posterior mean**: $\hat{\theta}^{\text{mean}} = \mathbb{E}[\theta|D]$.
  - **Posterior variance**: $\text{Var}[\theta|D]$. 
Bayesian Inference on the Categorical Distribution

Suppose we observe the with $y_i \in \{1, \ldots, K\}$, and model them as i.i.d. with pmf $\pi = (\pi_1, \ldots, \pi_K)$:

$$p(D|\pi) = \prod_{i=1}^{n} \pi_{y_i} = \prod_{k=1}^{K} \pi_k^{n_k}$$

with $n_k = \sum_{i=1}^{n} 1(y_i = k)$ and $\pi_k > 0$, $\sum_{k=1}^{K} \pi_k = 1$.

The conjugate prior on $\pi$ is the Dirichlet distribution $\text{Dir}(\alpha_1, \ldots, \alpha_K)$ with parameters $\alpha_k > 0$, and density

$$p(\pi) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

on the probability simplex $\{\pi : \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1\}$.

The posterior is also Dirichlet $\text{Dir}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$.

Posterior mean is

$$\hat{\pi}_k^{\text{mean}} = \frac{\alpha_k + n_k}{\sum_{j=1}^{K} \alpha_j + n_j}.$$
Dirichlet Distributions

(A) Support of the Dirichlet density for $K = 3$.
(B) Dirichlet density for $\alpha_k = 10$.
(C) Dirichlet density for $\alpha_k = 0.1$. 

Department of Statistics, Oxford
Naïve Bayes

Consider the classification example with naive Bayes classifier:

\[ p(x_i|\phi_k) = \prod_{j=1}^{p} \phi_{kj}^{x_{ij}} (1 - \phi_{kj})^{1-x_{ij}}. \]

Set \( n_k = \sum_{i=1}^{n} 1\{y_i = k\} \), \( n_{kj} = \sum_{i=1}^{n} 1\{y_i = k, x_{ij} = 1\} \). MLEs are:

\[ \hat{\pi}_k = \frac{n_k}{n}, \quad \hat{\phi}_{kj} = \frac{\sum_{i:y_i=k} x_{ij}}{n_k} = \frac{n_{kj}}{n_k}. \]

A problem: if the \( \ell \)-th word did not appear in documents labelled as class \( k \) then \( \hat{\phi}_{k\ell} = 0 \) and

\[ \mathbb{P}(Y = k|X = x \text{ with } \ell\text{-th entry equal to } 1) \]

\[ \propto \hat{\pi}_k \prod_{j=1}^{p} \left( \hat{\phi}_{kj} ight)^{x_{ij}} \left( 1 - \hat{\phi}_{kj} \right)^{1-x_{ij}} = 0 \]

i.e. we will never attribute a new document containing word \( \ell \) to class \( k \) (regardless of other words in it).
Bayesian Inference on Naïve Bayes model

- Under the Naïve Bayes model, the joint distribution of labels $y_i \in \{1, \ldots, K\}$ and data vectors $x_i \in \{0, 1\}^p$ is

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(x_i, y_i|\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left( \pi_k \prod_{j=1}^{p} \phi_{kj}^{x_{ij}} (1 - \phi_{kj})^{1-x_{ij}} \right)^{1(y_i=k)}$$

$$= \prod_{k=1}^{K} \pi_k^{n_k} \prod_{j=1}^{p} \phi_{kj}^{n_{kj}} (1 - \phi_{kj})^{n_k-n_{kj}}$$

where $n_k = \sum_{i=1}^{n} 1(y_i = k)$, $n_{kj} = \sum_{i=1}^{n} 1(y_i = k, x_{ij} = 1)$.

- For conjugate prior, we can use $\text{Dir}((\alpha_k)_k)$ for $\pi$, and $\text{Beta}(a, b)$ for $\phi_{kj}$ independently.

- Because the likelihood factorises, the posterior distribution over $\pi$ and $(\phi_{kj})$ also factorises, and posterior for $\pi$ is $\text{Dir}((\alpha_k + n_k)_k)$, and for $\phi_{kj}$ is $\text{Beta}(a + n_{kj}, b + n_k - n_{kj})$. 
Bayesian Inference on Naïve Bayes model

Given $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, want to predict a label $\tilde{y}$ for a new document $\tilde{x}$. We can calculate

$$p(\tilde{x}, \tilde{y} = k | \mathcal{D}) = p(\tilde{y} = k | \mathcal{D}) p(\tilde{x} | \tilde{y} = k, \mathcal{D})$$

with

$$p(\tilde{y} = k | \mathcal{D}) = \frac{\alpha_k + n_k}{\sum_{l=1}^{K} \alpha_l + n}, \quad p(\tilde{x}^{(j)} = 1 | \tilde{y} = k, \mathcal{D}) = \frac{a + n_{kj}}{a + b + n_k}.$$ 

Predicted class is

$$p(\tilde{y} = k | \tilde{x}, \mathcal{D}) = \frac{p(\tilde{y} = k | \mathcal{D}) p(\tilde{x} | \tilde{y} = k, \mathcal{D})}{p(\tilde{x} | \mathcal{D})} \propto \frac{\alpha_k + n_k}{\sum_{l=1}^{K} \alpha_l + n} \prod_{j=1}^{p} \left( \frac{a + n_{kj}}{a + b + n_k} \right)^{\tilde{x}^{(j)}} \left( \frac{b + n_k - n_{kj}}{a + b + n_k} \right)^{1-\tilde{x}^{(j)}}$$

Compared to ML plug-in estimator, pseudocounts help to “regularize” probabilities away from extreme values.
Bayesian Learning and Regularisation

- Consider a Bayesian approach to logistic regression: introduce a multivariate normal prior for weight vector $w \in \mathbb{R}^p$, and a uniform (improper) prior for offset $b \in \mathbb{R}$. The prior density is:

$$p(b, w) = 1 \cdot (2\pi\sigma^2)^{-\frac{p}{2}} \exp\left(-\frac{1}{2\sigma^2} \|w\|_2^2\right)$$

- The posterior is

$$p(b, w|D) \propto \exp\left(-\frac{1}{2\sigma^2} \|w\|_2^2 - \sum_{i=1}^{n} \log(1 + \exp(-y_i(b + w^\top x_i)))\right)$$

- The posterior mode is equivalent to minimising the $L_2$-regularised empirical risk.

- Regularised empirical risk minimisation is (often) equivalent to having a prior and finding a MAP estimate of the parameters.

  - $L_2$ regularisation - multivariate normal prior.
  - $L_1$ regularisation - multivariate Laplace prior.

- From a Bayesian perspective, the MAP parameters are just one way to summarise the posterior distribution.
A model $\mathcal{M}$ with a given set of parameters $\theta_{\mathcal{M}}$ consists of both the likelihood $p(D|\theta_{\mathcal{M}})$ and the prior distribution $p(\theta_{\mathcal{M}})$.

The posterior distribution

$$p(\theta_{\mathcal{M}}|D, \mathcal{M}) = \frac{p(D|\theta_{\mathcal{M}}, \mathcal{M})p(\theta_{\mathcal{M}}|\mathcal{M})}{p(D|\mathcal{M})}$$

Marginal probability of the data under $\mathcal{M}$ (Bayesian model evidence):

$$p(D|\mathcal{M}) = \int_{\Theta} p(D|\theta_{\mathcal{M}}, \mathcal{M})p(\theta_{\mathcal{M}}|\mathcal{M})d\theta$$

Compare models using their Bayes factors $\frac{p(D|\mathcal{M})}{p(D|\mathcal{M}')}$
Bayesian Occam’s Razor

- **Occam’s Razor**: of two explanations adequate to explain the same set of observations, the simpler should be preferred.

\[ p(D|M) = \int_{\Theta} p(D|\theta_M, M)p(\theta_M|M)d\theta \]

- Model evidence \( p(D|M) \) is the probability that a set of randomly selected parameter values inside the model would generate dataset \( D \).
- Models that are **too simple** are unlikely to generate the observed dataset.
- Models that are **too complex** can generate many possible dataset, so again, they are unlikely to generate that particular dataset at random.

![Graph showing the probability of datasets under different models](image-url)
Bayesian model comparison: Occam’s razor at work

figures by M. Sahani
Bayesian computation

Most posteriors are intractable, and posterior approximations need to be used.

- **Laplace approximation.**
- Variational methods (**variational Bayes**, expectation propagation).
- Monte Carlo methods (MCMC and SMC).
- Approximate Bayesian Computation (ABC).
Bayesian Learning – Discussion

- Use probability distributions to reason about uncertainties of parameters (latent variables and parameters are treated in the same way).
- Model consists of the likelihood function and the prior distribution on parameters: allows to integrate prior beliefs and domain knowledge.
- Prior usually has hyperparameters, i.e., \( p(\theta) = p(\theta|\psi) \). How to choose \( \psi \)?
  - Be Bayesian about \( \psi \) as well — choose a hyperprior \( p(\psi) \) and compute \( p(\psi|D) \): integrate the predictive posterior over hyperparameters.
  - Maximum Likelihood II — \( \hat{\psi} = \arg\max_{\psi \in \Psi} p(D|\psi) \).

\[
p(D|\psi) = \int p(D|\theta)p(\theta|\psi)d\theta
\]

\[
p(\psi|D) = \frac{p(D|\psi)p(\psi)}{p(D)}
\]
Bayesian Learning – Further Reading

- Videolectures by Zoubin Ghahramani: [Bayesian Learning](#)
- Murphy, Chapter 5