1. **(Clustering).** Let $x_1, \ldots, x_n$ be a dataset of $p$-dimensional vectors and $C = \{C_1, C_2, \ldots, C_K\}$ a partition of $\{1, \ldots, n\}$. For each cluster $C_k$, denote $n_k = |C_k|$ and define

$$
\bar{x}_k = \frac{1}{n_k} \sum_{i \in C_k} x_i
$$

to be the within-cluster mean

$$
\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k \bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_i
$$

to be the overall mean

and

$$
T = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x})(x_i - \bar{x})^\top
$$

to be the total deviance to the overall mean

$$
W = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^\top
$$

to be the within-cluster deviance to the cluster mean

$$
B = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^\top
$$

to be the between-cluster deviance

where $T$, $W$ and $B$ are all $p \times p$ matrices.

(a) Verify that $T = W + B$.

**Answer:**

$$
T = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x}_k + \bar{x}_k - \bar{x})(x_i - \bar{x}_k + \bar{x}_k - \bar{x})^\top
$$

$$
= \sum_{k=1}^{K} \sum_{i \in C_k} \left\{ (x_i - \bar{x}_k)(x_i - \bar{x}_k)^\top + (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^\top \right\} +
\sum_{k=1}^{K} \sum_{i \in C_k} \left\{ (x_i - \bar{x}_k)(\bar{x}_k - \bar{x})^\top + (\bar{x}_k - \bar{x})(x_i - \bar{x}_k)^\top \right\}
= W + B + 0 + 0
$$

since $\sum_{i \in C_k} (x_i - \bar{x}_k) = 0$, etc. The total variation is the sum of the within and between cluster variation.

The matrices $T$, $W$ and $B$ are proportional respectively to the total, within and between cluster sample covariance matrices.

(b) Explain how the K-means objective is related to the matrix $W$. Does it depend on all elements of $W$?

**Answer:** The K-means objective is exactly $Tr(W)$. 

1
(c) How does \(T\) change during the course of the K-means algorithm? How does \(B\) change?

**Answer:** \(Tr(T)\) does not depend on the partition so is constant. As K-means seeks to minimize \(Tr(W)\), it also equivalently seeks to increase \(Tr(B) = Tr(T) - Tr(W)\).

Thus just as PCA can be equivalently seen either as finding directions with maximum deviation or minimum reconstruction error, K-means can be seen equivalently either as minimizing within-cluster variance or as maximizing the between-cluster deviance.

2. (K-means). In lectures we discussed using the Mahalanobis distance to measure distances in K-means:

\[
\|x - y\|_M = \sqrt{(x - y)^\top M^{-1} (x - y)}
\]

where \(M\) is a positive definite matrix. Explain why using this distance is equivalent to applying K-means using the standard Euclidean distance on a transformed data set. What is the choice of the \(M\) matrix that leads to an algorithm which is equivalent to first whitening the data? [Hint: Consider a linear transformation \(x \mapsto Ax\).]

**Answer:** Let \(A = M^{-1/2}\) be the matrix square root of \(M^{-1}\), i.e., with eigendecomposition \(M = V \Lambda V^\top\), \(A = V \Lambda^{-1/2} V^\top\). Note \(A^\top = A\). Then

\[
\|x - y\|_M = \sqrt{(x - y)^\top M^{-1} (x - y)} = \sqrt{(x - y)^\top AA(x - y)}
\]

\[
= \sqrt{(Ax - Ay)^\top (Ax - Ay)} = \|Ax - Ay\|_2
\]

So the required transformation is \(x \mapsto Ax\) (\(X \mapsto XA\) for the overall data matrix).

In whitening the data, we seek a matrix \(A\) so that the sample covariance of \(\{Ax_i\}_{i=1}^n\) is \(I\). If \(S\) is the sample covariance, we can use \(A = S^{-1/2}\) so that the resulting sample covariance after transformation \(X \mapsto XA\) (assuming without loss of generality that mean is 0):

\[
\frac{1}{n-1} (XA)^\top (XA) = A \left( \frac{1}{n-1} X^\top X \right) A = ASA = I.
\]

Thus the choice is \(M = S\), the sample covariance.

3. (Deviance minimization equivalence). In the lectures, we considered within-cluster deviance \(W(C_k, \mu_k) = \sum_{i \in C_k} \|x_i - \mu_k\|^2\). Show that

\[
W(C_k, \mu_k) = W(C_k, \bar{x}_k) + |C_k| \cdot \|\mu_k - \bar{x}_k\|^2
\]

where \(\bar{x}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i\). Hence, conclude that \(W(C_k, \mu_k)\) is minimized for \(\mu_k = \bar{x}_k\) and show that

\[
W(C_k, \bar{x}_k) = \frac{1}{2|C_k|} \sum_{i,j \in C_k} \|x_i - x_j\|^2.
\]

**Answer:**
The first part follows from noticing
\[
\sum_{i \in C_k} \| x_i - \mu_k \|_2^2 = \sum_{i \in C_k} \| (x_i - \bar{x}_k) + (\bar{x}_k - \mu_k) \|_2^2 \\
= \sum_{i \in C_k} \left( \| x_i - \bar{x}_k \|_2^2 + \| \bar{x}_k - \mu_k \|_2^2 + 2 (x_i - \bar{x}_k)^T (\bar{x}_k - \mu_k) \right) \\
= \sum_{i \in C_k} \| x_i - \bar{x}_k \|_2^2 + |C_k| \cdot \| \bar{x}_k - \mu_k \|_2^2 + 2 \left( \sum_{i \in C_k} x_i - |C_k| \bar{x}_k \right)^T (\bar{x}_k - \mu_k).
\]

For the second equation, we proceed similarly, starting from the right hand side:

\[
\frac{1}{|C_k|} \sum_{i,j \in C_k} \| x_i - x_j \|_2^2 = \frac{1}{|C_k|} \sum_{i,j \in C_k} \| (x_i - \bar{x}_k) - (x_j - \bar{x}_k) \|_2^2 \\
= \frac{1}{|C_k|} \sum_{i,j \in C_k} \left( \| x_i - \bar{x}_k \|_2^2 + \| x_j - \bar{x}_k \|_2^2 - 2 (x_i - \bar{x}_k)^T (x_j - \bar{x}_k) \right) \\
= 2 \sum_{i \in C_k} \| x_i - \bar{x}_k \|_2^2 - \frac{2}{|C_k|} \sum_{j \in C_k} \left( \sum_{i \in C_k} x_i - |C_k| \bar{x}_k \right)^T (x_j - \bar{x}_k) \\
= 2W(C_k, \bar{x}_k).
\]

4. **(Regression loss, squared and absolute).** For a given loss function \( L \), the risk \( R \) is given by the expected loss

\[
R(f) = \mathbb{E} [L(Y, f(X))],
\]

where \( f = f(X) \) is a function of the random predictor variable \( X \).

(a) Consider a regression problem and the squared error loss

\[
L(Y, f(X)) = (Y - f(X))^2.
\]

Derive the expression of \( f = f(X) \) minimizing the associated risk.

**Answer:** We have

\[
R(f) = \mathbb{E} \left[ (Y - f(X))^2 \right] = \int \mathbb{E} \left[ (Y - f(X))^2 \bigg| X = x \right] g_X(x) \, dx,
\]

where \( g_X \) is density of \( X \). Thus, it suffices to for every \( x \), minimize:

\[
\mathbb{E} \left[ (Y - f(X))^2 \bigg| X = x \right] = \mathbb{E} [Y^2 | X = x] - 2 f(x) \mathbb{E} [Y | X = x] + f(x)^2 = \text{Var}[Y | X = x] + (\mathbb{E} [Y | X = x] - f(x))^2.
\]

This is clearly minimized by the conditional mean:

\[
f(x) = \mathbb{E} [Y | X = x].
\]
(b) What if we use the absolute ($L_1$) loss instead?

$$L(Y, f(X)) = |Y - f(X)|.$$ 

**Answer:**

As before, we want to find $f(x)$ to minimize

$$\mathbb{E} [ |Y - f(x)| | X = x ] = \int_{-\infty}^{\infty} |y - f(x)| \ p(y) \ dy$$

$$= \int_{-\infty}^{f(x)} (f(x) - y) \ p(y) \ dy + \int_{f(x)}^{\infty} (y - f(x)) \ p(y) \ dy$$

We differentiate these integrals with respect to $f(x)$ using the Leibniz integral rule:

$$\frac{d}{dx} \left( \int_{a(x)}^{b(x)} g(x, t) \ dt \right) = g(x, b(x)) \cdot \frac{d}{dx} b(x) - g(x, a(x)) \cdot \frac{d}{dx} a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} g(x, t) \ dt,$$

which gives

$$\frac{d}{df(x)} \mathbb{E} [ |Y - f(x)| | X = x ] = \int_{-\infty}^{f(x)} p(y) \ dy + \int_{f(x)}^{\infty} -p(y) \ dy$$

$$= P(Y < f(x)|X = x) - P(Y > f(x)|X = x).$$

Setting this to 0, we get the minimum for $P(Y < f(x)|X = x) = P(Y > f(x)|X = x) = 1/2$, i.e. at the conditional median $f(x) = \text{med} \{ Y | X = x \}$.

5. **(ESL: Train/test)** Consider a linear regression model with $p$ parameters, fit with unregularized linear regression (sometimes called “least squares” or “ordinary least squares” or even just OLS) to a set of training data $(x_i, y_i)_{1 \leq i \leq N}$ drawn at random from a population. Let $\hat{\beta}$ be the estimator. Suppose we have some test data $(\tilde{x}_i, \tilde{y}_i)_{1 \leq i \leq M}$ drawn at random from the same population as the training data.

If $R_{tr}(\beta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$ and $R_{te}(\beta) = \frac{1}{M} \sum_{i=1}^{M} (\tilde{y}_i - \tilde{x}_i^T \beta)^2$, prove that

$$E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})]$$

where the expectation is over all that is random in each expression.

**Answer:**

Assume we have two sets of $N$ i.i.d. data points from the same distribution $P(X,Y)$. Call these $\{(x_i, y_i)\}_{i=1}^{N}$ and $\{((\tilde{x}_i, \tilde{y}_i))\}_{i=1}^{N}$. We will initially ignore the fact that we have $M$ testing points, and address this later. This will allow us to work with estimators built on training data sets of the same size, which have the same expected out-of-sample (testing) error. Define the OLS estimators as

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2$$

And

$$\tilde{\beta} = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} (\tilde{y}_i - \beta^T \tilde{x}_i)^2$$
We can write

\[ E[R_{tr}(\hat{\beta})] = E \left[ \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\beta}^T x_i)^2 \right] \]
\[ = \frac{1}{N} \sum_{i=1}^{N} E \left[ (y_i - \hat{\beta}^T x_i)^2 \right] \]

From the definition of \( \hat{\beta} \), it follows that \( E \left[ (y_i - \hat{\beta}^T x_i)^2 \right] \leq E \left[ (y_i - \beta^T x_i)^2 \right], \forall \beta \), where \( \{(x_i, y_i)\}_{i=1}^{N} \) are the points where \( \hat{\beta} \) was trained on. Because this is true for all \( \beta \), it is also true for \( \tilde{\beta} \), so we can write:

\[ \frac{1}{N} \sum_{i=1}^{N} E \left[ (y_i - \hat{\beta}^T x_i)^2 \right] \leq \frac{1}{N} \sum_{i=1}^{N} E \left[ (y_i - \tilde{\beta}^T x_i)^2 \right] \]

But because points in each data set are i.i.d., and the definitions of \( \hat{\beta} \) and \( \tilde{\beta} \) are symmetric, we have that

\[ E[R_{tr}(\hat{\beta})] \leq \frac{1}{N} \sum_{i=1}^{N} E \left[ (y_i - \hat{\beta}^T x_i)^2 \right] \]
\[ = \frac{1}{N} \sum_{i=1}^{N} E \left[ (\tilde{y}_i - \hat{\beta}^T \tilde{x}_i)^2 \right] \]
\[ = E \left[ (\tilde{y}_i - \hat{\beta}^T \tilde{x}_i)^2 \right] \]
\[ = \frac{1}{M} \sum_{i=1}^{M} E \left[ (\tilde{y}_i - \hat{\beta}^T \tilde{x}_i)^2 \right] \]
\[ = E \left[ R_{te}(\hat{\beta}) \right], \]

where have taken into account that the test set is made of \( M \) data points. Thus,

\[ E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})] \]

6. (Weighted loss). Consider two univariate normal distributions \( N(\mu, \sigma^2) \) with known parameters \( \mu_A = 10 \) and \( \sigma_A = 5 \) for class A and \( \mu_B = 20 \) and \( \sigma_B = 5 \) for class B. Suppose class A represents the random score \( X \) of a medical test of normal patients and class B represents the score of patients with a certain disease. A priori there are 100 times more healthy patients than patients carrying the disease.
(a) Find the optimal decision rule in terms of misclassification error (0-1 loss) for allocating a new observation \( x \) to either class A or B.

**Answer:** The optimal decision for \( X = x \) is to allocate to class

\[
\text{argmax}_{k \in \{A, B\}} \pi_k g_k(x).
\]

The patients should be classified as healthy iff

\[
\pi_A \frac{1}{\sqrt{2\pi \sigma_A}} \exp \left( -\frac{(x - \mu_A)^2}{2\sigma_A^2} \right) \geq \pi_B \frac{1}{\sqrt{2\pi \sigma_B}} \exp \left( -\frac{(x - \mu_B)^2}{2\sigma_B^2} \right),
\]

that is, using \( \sigma_A = \sigma_B \), iff

\[
-2\sigma_A^2 \log(\pi_A/\pi_B) + (x - \mu_A)^2 \leq (x - \mu_B)^2.
\]

The decision boundary is attained for equality, that is if \( x \) fulfills

\[
2x(\mu_B - \mu_A) + \mu_A^2 - \mu_B^2 - 2\sigma_A^2 \log(\pi_A/\pi_B) = 0.
\]

For the given values, this implies that the decision boundary is at

\[
x = \frac{(50 \log 100 - 100 + 400)}{2 \cdot 10} \approx 26.51,
\]

that is all patients with a test score above 26.51 are classified as having the disease.

(b) Repeat (a) if the cost of a false negative (allocating a sick patient to group A) is \( \theta > 1 \) times that of a false positive (allocating a healthy person to group B). Describe how the rule changes as \( \theta \) increases. For which value of \( \theta \) are 84.1% of all patients with disease correctly classified?

**Answer:** The optimal decision minimizes \( E[ L(Y, f(x)) | X = x] \). The risk for choosing A is now

\[
L(A, A) P(Y = A | X = x) + L(B, A) P(Y = B | X = x) = 0 + \theta P(Y = B | X = x),
\]

and the risk for B is

\[
L(A, B) P(Y = A | X = x) + L(B, B) P(Y = B | X = x) = P(Y = A | X = x) + 0.
\]

It is hence optimal to choose class A (healthy) over class B if and only if

\[
P(Y = A | X = x) \geq \theta P(Y = B | X = x).
\]

Using the same argument as above, the patients should be classified as healthy now iff (ignoring again the common denominator \( \sum_{k \in \{A, B\}} \pi_k g_k(x) \)),

\[
\pi_A \frac{1}{\sqrt{2\pi \sigma_A}} \exp \left( -\frac{(x - \mu_A)^2}{2\sigma_A^2} \right) \geq \theta \pi_B \frac{1}{\sqrt{2\pi \sigma_B}} \exp \left( -\frac{(x - \mu_B)^2}{2\sigma_B^2} \right),
\]

The decision boundary is now attained if \( x \) fulfills

\[
2x(\mu_B - \mu_A) + \mu_A^2 - \mu_B^2 - 2\sigma_A^2 \log(\pi_A/(\theta \pi_B)) = 0.
\]

For increasing values of \( \theta \), patients with smaller test scores are classified as having the disease (e.g. further screened).
84.1% of all patients carrying the disease are correctly classified if the decision boundary is at the 15.9%-quantile of the $\mathcal{N}(\mu_B, \sigma_B^2)$-distribution, which is at $q = 20 + 5\Phi^{-1}(0.159) \approx 15$. This decision boundary is attained if

$$15 = q = (50 \log(100/\theta) - 100 + 400)/20,$$

which implies that for

$$\theta = 100 \exp \left( -\frac{20q - 300}{50} \right) = 100,$$

approximately 84.1% of all patients are correctly classified as carrying the disease.

7. (*Coding: MDS dissimilarity measure, hierarchical clustering*). Download cognate.txt from http://www.stats.ox.ac.uk/~palamara/teaching/SML19/cognate.txt and load it using

```r
X <- read.table("cognate.txt")
```

It contains an $87 \times 2665$ matrix of observations on each of 87 Indo-European languages where the presence (1) or absence (0) of 2665 homologous traits has been recorded.

Historical linguists have grouped these languages into clades. Most large-scale groupings are contested, but something like

$$\{\text{Indic, Iranian}\}$$

$$\{\text{Balto–Slav, (Germanic, Italic, Celtic)}\}$$

is not too controversial. The position of the Armenian, Greek, Albanian, Tocharian and Hittite groups is in doubt (though not within the second of the above super-clade).

We would like to cluster the languages into groups on the basis of these data. It is also of interest to represent the languages in a planar map in order to visualise similarities between languages.

(a) These data are categorical. The Simple Matching Coefficient for two data vectors is the proportion of variables which are unequal. The Jaccard coefficient for two language data vectors is the proportion of variables with at least one present which are unequal (so 1100 and 1010 have SMC 2/4 and JC 2/3). Which dissimilarity measure is appropriate for these data and why?

**Answer:** Probably Jaccard. These are trait data. If two objects both lack many irrelevant traits, that should not make them more similar. So dist(1100,1000) and d(1111,1100) should be the same, a vote for Jaccard. In this data 1’s make up only about 5% of the data values, so shared absence of traits is much more common than shared presence, again pointing to Jaccard.

(b) Run MDS with Sammon mapping using both SMC and Jaccard distance on these data. You can use

```r
D<-dist(X,method="binary")
```

for computing the Jaccard distances, and

```r
D<-dist(X,method="manhattan")
```

for SMC.

(c) Compute agglomerative clustering of the data using Jaccard with single, average and complete linkage. Plotting the dendrograms with language labels on the leaves, which linkage algorithm seems to produce sensible results? You can use

```r
hclust(D,method=...) or agnes(D,method=...) for various choices of linkage.
```

(agnes is part of the cluster library, so you have to load using `library(cluster)`).
Answer: Using Jaccard and average linkage seem to produce the most sensible dendrogram. This reproduces most of the prior clade structure. complete is similar while single produces strange groupings.

X <- read.table("cognate.txt")
## calculate Jaccard distances
D <- dist(X,method="binary")
## form clusters by agglomerative cluster using "average"
## or "complete". The choice "single" is not robust.
hi <- hclust(D, method="average")
plot(hi,labels=row.names(X),cex=0.6,font=2,ann=FALSE)

8. (Optional: ESL Bayesian L2). (Assumes some knowledge of Bayesian statistics.) We are performing linear regression using the squared loss, but we place a Gaussian prior on the vector of coefficients we are trying to learn: $\beta_j \sim N(0, \sigma_\beta^2)$, for $j \in \{1, \ldots, p\}$. We assume that the observations $y_i$ are sampled adding Gaussian noise with variance $\sigma_\epsilon^2$ to points from the underlying linear model given by $\beta_0 + x_i^T \beta$, so that $y_i \sim N(\beta_0 + x_i^T \beta, \sigma_\epsilon^2)$, for $i \in \{1, \ldots, n\}$. Find the maximum-a-posteriori (which is also the posterior mean) for the vector of coefficients $\beta$. Show the equivalence between this estimator and the estimator obtained when performing L2-regularized (ridge) regression, and interpret the relationship between the regularization parameter $\lambda$ and the parameters $\sigma_\beta^2$ and $\sigma_\epsilon^2$. [Hint: marginal and conditional distributions of multivariate Gaussians have standard expressions].

Answer: To simplify notation we will write $X \beta$ for the linear model, but note that we are not placing a prior on $\beta_0$, as this would shrink the intercept to 0 and penalize shifting the data by a fixed amount, which we don’t need to do. We have

$$p(\beta) = N(0, \sigma_\beta^2 I_p)$$

And

$$p(y|\beta) = N(X \beta, \sigma_\epsilon^2 I_n)$$

We are interested in the mean of the posterior $p(\beta|y)$. We can multiply $p(\beta)$ and $p(y|\beta)$ to get the joint distribution. After some algebra and manipulations, we find that the posterior $p(\beta|y)$ is also
normally distributed, with mean
\[ m = (\sigma_{\beta}^{-2} I_p + \sigma_{\epsilon}^{-2} X^T X)^{-1} \sigma_{\epsilon}^{-2} X^T y = (\lambda I_p + X^T X)^{-1} X^T y, \]

where we have set \( \lambda = \sigma_{\epsilon}^2 / \sigma_{\beta}^2 \). Because the mean is also the maximum of this distribution, this will correspond to the maximum-a-posteriori (MAP) estimate of \( \hat{\beta} \). So the larger the ratio between our assumed variance and the observation noise, the weaker the regularization, and the induced shrinkage.