SC4/SM8 Advanced Topics in Statistical Machine Learning Problem Sheet 1

1. Suppose we do PCA, projecting each x_i into $z_i = V_{1:k}^\top x_i$ where $V_{1:k} = [v_1, \dots, v_k]$, i.e., the first k principal components. We can reconstruct x_i from z_i as $\hat{x}_i = V_{1:k} z_i$.

   (a) Show that $\|\hat{x}_i - \hat{x}_j\|_2 = \|z_i - z_j\|_2$.

   (b) Show that the error in the reconstruction equals:

   $$\sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2 = (n-1) \sum_{j=k+1}^{p} \lambda_j$$

   where $\lambda_{k+1}, \dots, \lambda_p$ are the $p-k$ smallest eigenvalues. Thus, the more principal components we use for the reconstruction, the more accurate it is. Further, using the top k principal components is optimal in the sense of least reconstruction error.

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

   $$\bar{x}_k = \frac{1}{n_k} \sum_{i \in C_k} x_i \qquad \text{to be the within-cluster mean}$$

   $$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k \bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{to be the overall mean}$$

   and

   $$T = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x})(x_i - \bar{x})^\top \qquad \text{to be the \textit{total deviance matrix}, i.e. to the overall mean}$$

   $$W = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^\top \qquad \text{to be the \textit{within-cluster deviance matrix}, i.e. to the cluster means}$$

   $$B = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^\top \qquad \text{to be the \textit{between-cluster deviance matrix}}$$

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

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

   (b) Explain how the K-means objective is related to W.

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

3. For a given loss function L, the risk R is given by the expected loss

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

   where we consider real-valued responses, i.e. $f : \mathcal{X} \to \mathbb{R}$. Derive the optimal regression functions (which minimize the associated risk) for the following losses:

   (a) The squared error loss

   $$L(Y, f(X)) = (Y - f(X))^2$$

(b) The absolute (L_1) loss

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

(c) The τ-pinball loss, $\tau \in (0, 1)$

$$L(Y, f(X)) = 2 \max \{\tau(Y - f(X)), (\tau - 1)(Y - f(X))\}$$

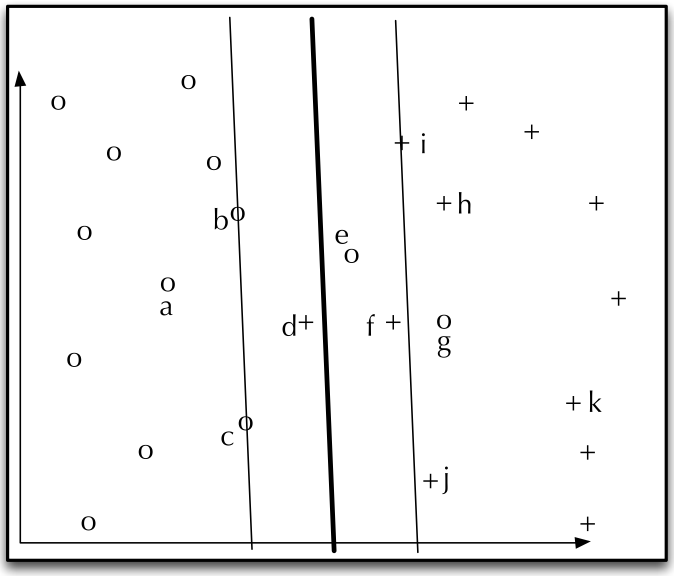
4. In binary classification, suppose that $\mathbb{P}(Y = -1)$ is very small, so that the constant classifier $f(x) = +1, \forall x$, has a small risk under the 0/1 loss. Consider the following loss instead:

$$L_{\alpha,\beta}(Y, f(X)) = \begin{cases} \alpha & \text{if } Y = -1, f(X) = +1, \\ \beta & \text{if } Y = +1, f(X) = -1, \\ 0 & \text{otherwise.} \end{cases}$$

Find α and β that result in the following risk

$$R(f) = \mathbb{P}(f(X) = +1|Y = -1) + \mathbb{P}(f(X) = -1|Y = +1).$$

5. The figure below shows a binary classification dataset and the optimal the decision boundary and margins of a soft-margin C-SVM for some value C.



(i) Which of the points a,...,k are support vectors? Which ones are margin support vectors?

(ii) For points a, b and d what are the range of possible values for the corresponding dual variables?

6. Parameter C in C-SVM can sometimes be hard to interpret. An alternative parametrization is given by ν-SVM:

$$\min_{w,b,\rho,\xi} \left(\frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n} \xi_i \right)$$

subject to

$$\begin{aligned} \rho &\geq 0, \\ \xi_i &\geq 0, \\ y_i\left(w^\top x_i + b\right) &\geq \rho - \xi_i. \end{aligned}$$

(note that we now directly adjust the constraint threshold ρ).

Using complementary slackness, show that ν is an upper bound on the proportion of non-margin support vectors (margin errors) and a lower bound on the proportion of all support vectors with non-zero weight (both those on the margin and margin errors). You can assume that $\rho > 0$ at the optimum (non-zero margin).

7. **(Kernel Ridge Regression)** Let $(x_i, y_i)_{i=1}^n$ be our dataset, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Classical linear regression can be formulated as empirical risk minimization, where the model is to predict y using a class of functions $f(x) = w^\top x$, for some vector $w \in \mathbb{R}^p$ and we use the squared loss, i.e. we minimize

$$R^{\text{emp}}(w) = \frac{1}{n}\sum_{i=1}^n (y_i - w^\top x_i)^2.$$

(a) Show that the optimal parameter is

$$\hat{w} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$$

where $\mathbf{X}$ is a $n \times p$ matrix with ith row given $x_i^\top$, and $\mathbf{y}$ is a $n \times 1$ matrix with ith entry y_i.

(b) Consider regularizing our empirical risk by incorporating an L_2 regularizer. That is, find w minimizing

$$\frac{1}{n}\sum_{i=1}^n (y_i - w^\top x_i)^2 + \frac{\lambda}{n}\|w\|_2^2$$

Show that the optimal parameter is given by the ridge regression estimator

$$\hat{w} = (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top \mathbf{y}$$

(c) Suppose that we now wish to introduce nonlinearities into the model, by transforming $x \mapsto \varphi(x)$. Show how this transformation may be achieved using the kernel trick. That is, let $\mathbf{\Phi}$ be a matrix with ith row given by $\varphi(x_i)^\top$. The optimal parameters $\hat{w}$ would then be given by (previous part):

$$\hat{w} = (\mathbf{\Phi}^\top \mathbf{\Phi} + \lambda I)^{-1}\mathbf{\Phi}^\top \mathbf{y}.$$

Can we make predictions without computing $\hat{w}$?

First, express the predicted y values on the training set, $\mathbf{\Phi}\hat{w}$, only in terms of $\mathbf{y}$ and the Gram matrix $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^\top$, with $\mathbf{K}_{ij} = \varphi(x_i)^\top \varphi(x_j) = k(x_i, x_j)$ where k is some kernel function. Then, compute an expression for the value of $y_\star$ predicted by the model at an unseen test vector $x_\star$.

[*Hint: You will find the Woodbury matrix inversion formula useful:*

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

where A and B are square invertible matrices of size $n \times n$ and $p \times p$ respectively, and U and V are $n \times p$ and $p \times n$ rectangular matrices.]