Statistical Machine Learning HT 2018 - Problem Sheet 3

1. **LDA and QDA.** Suppose we have a two-class setup with classes -1 and 1, i.e., $\mathcal{Y} = \{-1, 1\}$, and a 2-dimensional predictor variable X. We find that the means of the two groups are at $\hat{\mu}_{-1} = (-1, -1)^\top$ and $\hat{\mu}_1 = (1, 1)^\top$ respectively. The estimated prior class probabilities $\hat{\pi}_1$ and $\hat{\pi}_{-1}$ are equal.

   (a) Applying LDA, the covariance matrix is estimated to be, for some value of $0 \leq \rho \leq 1$,

   $$\hat{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

   Find the decision boundary as a function of ρ.

   (b) Suppose instead that, we model each class with its own covariance matrix. We estimate the covariance matrices for group -1 as

   $$\hat{\Sigma}_{-1} = \begin{pmatrix} 5 & 0 \\ 0 & 1/5 \end{pmatrix},$$

   and for group 1 as

   $$\hat{\Sigma}_1 = \begin{pmatrix} 1/5 & 0 \\ 0 & 5 \end{pmatrix}.$$

   Describe the decision rule and draw a sketch of it in the two-dimensional plane.

2. **Naive Bayes vs Logistic regression.** The binary Naive Bayes classifier has interesting connections to the logistic regression classifier. You will show that, under certain assumptions, the Naive Bayes likelihood function is identical in form to the likelihood function for logistic regression. You will then derive the MLE parameter estimates under these assumptions.

   (a) Suppose $X = \{X_1, \ldots, X_D\}$ is a continuous vector in $\mathbb{R}^D$ representing the features, and Y is a binary random variable with values in $\{0, 1\}$ representing the class labels. Let the following assumptions hold:

   - The label variable Y follows a Bernoulli distribution, with parameter $\pi = P(Y = 1)$.
   - For each feature X_j, we have $P(X_j|Y = k)$ follows a Gaussian distribution of the form $\mathcal{N}(\mu_{jk}, \sigma_j)$.

   Using the Naive Bayes assumption that states *"for all $j' \neq j$, X_j and $X_{j'}$ are conditionally independent given Y"*, compute $P(Y = 1|X)$ and show that it can be written in the following form:

   $$P(Y = 1|X) = \frac{1}{1 + \exp(-w_0 + \mathbf{w}^\top \mathbf{X})}.$$

   Specifically, you need to find the explicit form of w_o and $\mathbf{w}$ in terms of π, μ_{jk}, σ_j, for $j = 1, \ldots, D$ and $k \in \{0, 1\}$.

   (b) Suppose a training set with N examples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$ is given, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})^\top$ is a D-dimensional feature vector, and $y_i \in \{0, 1\}$ is its corresponding label. Using the assumptions in 1.a (not the result), provide the maximum likelihood estimation for the parameters of the Naive Bayes with Gaussian assumption. In other words, you need to provide the estimates for π, μ_{jk}, and σ_j, for $j = 1, \ldots, D$ and $k \in \{0, 1\}$.

3. **Missing data in generative/discriminative models.** Assume we trained a classifier using data $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ where $x_i \in \mathbb{R}^p$ and $y_i \in \{1, 2, ..., K\}$. We are interested in classifying a new input vector $\tilde{x}$. However, we have only been able to collect $p-1$ features, say $\left(\tilde{x}^{(2)}, ..., \tilde{x}^{(p)}\right)$ and $\tilde{x}^{(1)}$ is missing. Explain whether or not it is possible to use the trained classifier to classify this incomplete input vector in the cases listed below. If it is possible, how do you classify the incomplete test vector?

   Note: You do not need to calculate any integrals in this question.

   (a) A naïve Bayes model, with
   $$g_k(x) = \prod_{j=1}^{p} p(x^{(j)}|\phi_{kj}),$$
   i.e. conditioned upon $Y = k$, you assume that the features are independent and feature $x^{(j)}$ has probability mass function/density $p(x^{(j)}|\phi_{kj})$.

   (b) An LDA model, i.e.
   $$g_k(x) = \mathcal{N}(x; \mu_k, \Sigma)$$

   (c) Generally, what condition on the conditional density/pmf $g_k(x)$ would allow easy classification (i.e, without numerical integration) in the presence of missing features for generative classifiers like LDA or naïve Bayes?

   (d) A logistic regression model, i.e.
   $$p(Y = y|X = x) = s(y(a + b^\top x))$$
   where $y \in \{+1, -1\}$.

4. **KNN and the curse of dimensionality.** Consider using a k-NN classifier where the real-valued features are uniformly distributed in the p-dimensional unit cube. Suppose we are interested in estimating the distribution over class labels around a test point x by using neighbours within a hyper-cube centred at x.

   (a) Suppose we wish to use a fraction α of the training data to estimate the distribution over class labels at x. What should be the edge length of this hyper-cube to ensure that it includes on average $\alpha\%$ of the training data? If $p = 10$ and $\alpha = 1\%$, compute the edge length of this hyper-cube. In this scenario, is k-NN a "local" algorithm, i.e. using only local neighbours to x?

   (b) Assuming you have access to say $n = 500$ training data (and $p = 10$ as before), does it appear reasonable to perform k-NN for large values of k (say $k > 10$)? Explain briefly why or why not.

5. **ROC.** The receiver operating characteristic (ROC) curve plots the sensitivity against the specificity of a binary classifier as the threshold for discrimination is varied.

   Let the data space be $\mathbb{R}$, and denote the class-conditional densities with $g_0(x)$ and $g_1(x)$ for $x \in \mathbb{R}$ and for the two classes 0 and 1. Consider a classifier that classifies x as class 1 if $x \geq c$, where threshold c varies from $-\infty$ to $+\infty$.

   (a) Give expressions for the (population versions of) specificity and sensitivity of this classifier.

   (b) Show that the AUC corresponds to the probability that $X_1 > X_0$, where data items X_1 and X_0 are independent and come from classes 1 and 0 respectively.

6. ***Coding*: LDA.** Download
`http://www.stats.ox.ac.uk/~palamara/teaching/SML18/wine.data`
and load it using `read.table("wine.data",sep=",")`. Description of the dataset is given at `https://archive.ics.uci.edu/ml/datasets/Wine`. The goal is to build a classifier for predicting the cultivars given in column 1. Train LDA classifier on a random subset of 50% of the data, and show the projections of the data vectors as well as the decision boundaries in the 2D LDA component space. Then predict the cultivars for the other 50% of the data and plot these in the LDA component space as well (using a different `pch`). How many errors did the classifier make (a) on the training set, (b) on the test set?

7. ***Coding*: logistic regression (and KNN).** (Exercise 2.8 in *Elements of Statistical Learning*) Compare the classification performance of logistic regression, regularized logistic regression (and optionally k-nearest neighbor classification) on the ZIP code digit image dataset, restricting to only the 2s and 3s. Investigate L1 and L2 regularization alone and in combination (the "elastic net"). Show both training and testing error for each choice. The ZIP code data are available from `https://web.stanford.edu/~hastie/ElemStatLearn/data.html`. You can read more about the data in Section 11.7 of *Elements of Statistical Learning*.

8. ***Optional*: 1-NN risk in binary classification.** Let $\{(X_i, Y_i)\}_{i=1}^n$ be a training dataset where $X_i \in \mathbb{R}^p$ and $Y_i \in \{0, 1\}$. We denote by $g_k(x)$ the conditional density of X given $Y = k$ and assume that $g_k(x) > 0$ for all $x \in \mathbb{R}^p$, and the class probabilities as $\pi_k = \mathbb{P}(Y = k)$. We further denote $q(x) = \mathbb{P}(Y = 1|X = x)$.

   (a) Consider the Bayes classifier (minimizing risk w.r.t. 0/1 loss $\mathbf{1}\{f(X) \neq Y\}$):

   $$f_{\text{Bayes}}(x) = \underset{k\in\{0,1\}}{\arg\max}\, \pi_k g_k(x).$$

   Write the conditional expected loss $\mathbb{P}[f(X) \neq Y|X = x]$ at a given test point $X = x$ in terms of $q(x)$. [The resulting expression should depend *only* on $q(x)$].

   (b) The 1-nearest neighbour (1-NN) classifier assigns to a test data point x the label of the closest training point; i.e. $f_{1\text{NN}}(x) = y$ (class of nearest neighbour in the training set). Given some test point $X = x$ and its nearest neighbour $X' = x'$, what is the conditional expected loss $\mathbb{P}[f_{1\text{NN}}(X) \neq Y|X = x, X' = x']$ of the 1-NN classifier in terms of $q(x)$, $q(x')$?

   (c) As the number of training examples goes to infinity, i.e. $n \to \infty$, assume that the training data fills the space such that $q(x') \to q(x)$, $\forall x$. Give the limit (as $n \to \infty$) of $\mathbb{P}[f_{1\text{NN}}(X) \neq Y|X = x]$. If we denote by $R_{\text{Bayes}} = \mathbb{P}\left[Y \neq f_{\text{Bayes}}(X)\right]$ and $R_{1\text{NN}} = \mathbb{P}[Y \neq f_{1\text{NN}}(X)]$, show that for sufficiently large n

   $$R_{\text{Bayes}} \leq R_{1\text{NN}} \leq 2R_{\text{Bayes}}\left(1 - R_{\text{Bayes}}\right).$$