ORF 525: Statistical Foundations of Data Science

Jianqing Fan — Frederick L. Moore'18 Professor of Finance

Problem Set #4 Spring 2021

Due Sunday, April 4, 2021.

Choose any of the 5 problems

1. Suppose that random variables representing two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ follow p-variate normal distributions $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively. Consider the linear classifier $\delta(\mathbf{X}) = I(\mathbf{w}^T(\mathbf{X} - \boldsymbol{\mu}) > 0) + 1$ which classifies $\mathbf{X}$ to class 2 when $\mathbf{w}^T(\mathbf{X} - \boldsymbol{\mu}) > 0$ for given parameters $\boldsymbol{\mu}$ and $\mathbf{w}$.

   (a) The misclassification rate of classifying a data point from class $\mathcal{C}_2$ is $P_2(\mathbf{w}^T(\mathbf{X} - \boldsymbol{\mu}) \leq 0)$ where P_2 is the probability distribution under $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Show that this misclassification rate is given by

   $$1 - \Phi\Big(\frac{\mathbf{w}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu})}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{1/2}}\Big).$$

   (b) The misclassification rate of classifying a data point from class $\mathcal{C}_1$ is $P_1(\mathbf{w}^T(\mathbf{X} - \boldsymbol{\mu}) > 0)$ where P_1 is the probability distribution under $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$. Show that this misclassification rate is given by

   $$1 - \Phi\Big(\frac{\mathbf{w}^T(\boldsymbol{\mu} - \boldsymbol{\mu}_1)}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{1/2}}\Big).$$

   (c) Define the prior probability $\pi = P(Y = 2)$. Show that the misclassification rate is

   $$P(\delta(\mathbf{X}) \neq Y) = 1 - \pi\Phi\Big(\frac{\mathbf{w}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu})}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{1/2}}\Big) - (1 - \pi)\Phi\Big(\frac{\mathbf{w}^T(\boldsymbol{\mu} - \boldsymbol{\mu}_1)}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{1/2}}\Big).$$

2. Consider the kernel density estimator $\widehat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n h^{-p}K((\mathbf{x}_i - \mathbf{x})/h)$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. random variables in $\mathbb{R}^p$ with density $f(\cdot)$ and $\mathbf{x} \in \mathbb{R}^p$, $h \in \mathbb{R}$.

   (a) For any $\mathbf{x}$ and h, show that $E\widehat{f}(\mathbf{x}) = \int K(\mathbf{y})f(\mathbf{x} + h\mathbf{y})d\mathbf{y}$.

   (b) If $\int y_i K(\mathbf{y})d\mathbf{y} = 0$ for all i and $f''(\cdot)$ is continuous at point $\mathbf{x}$, then

   $$E\widehat{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{h^2}{2}\int \mathbf{y}^T f''(\mathbf{x})\mathbf{y}K(\mathbf{y})d\mathbf{y} + o(h^2).$$

   If further $\int y_i y_j K(\mathbf{y})d\mathbf{y} = 0$ for all $i \neq j$, show that the bias

   $$E\widehat{f}(\mathbf{x}) - f(\mathbf{x}) = \frac{h^2}{2}\sum_{j=1}^p \frac{\partial^2}{\partial x_j^2} f(\mathbf{x}) \int y_j^2 K(\mathbf{y})d\mathbf{y} + o(h^2)$$

   Hint: Use Taylor expansion of $f(\mathbf{x} + h\mathbf{y})$. You may assume that the support of $K(\cdot)$ is bounded so that the limit and integral are exchangeable.

   (c) Show that the variance

   $$\mathrm{var}(\widehat{f}(\mathbf{x})) = \frac{1}{nh^p} f(\mathbf{x}) \int K^2(\mathbf{y})d\mathbf{y}(1 + o(1)).$$

3. Let Y be a random variable, taking values in $\{1, -1\}$. Let $p_+(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$ and $p_-(\mathbf{x}) = P(Y = -1|\mathbf{X} = \mathbf{x})$.

   (a) Show that
   $$\mathrm{argmin}_f \, \mathrm{E}\left([1 - Yf(\mathbf{X})]_+\right) = \mathrm{sign}(p_+(\mathbf{x}) - p_-(\mathbf{x})),$$
   i.e. the function $\mathrm{sign}(p_+(\mathbf{x}) - p_-(\mathbf{x}))$ achieves the minimum of $\mathrm{E}\left([1 - Yf(\mathbf{X})]_+\right)$.

   (b) Show that
   $$\mathrm{argmin}_f \, \mathrm{E}\log(1 + e^{-Yf(\mathbf{X})}) = \log\left(\frac{p_+(\mathbf{x})}{p_-(\mathbf{x})}\right).$$

   (c) Let ϕ be a convex loss function and if it is differentiable at 0 with $\phi'(0) < 0$, then ϕ is Fisher-consistent, namely
   $$\mathrm{sign}(f^*(\mathbf{x})) = \mathrm{sign}(p_+(\mathbf{x}) - p_-(\mathbf{x})), \quad \text{where} \quad f^* = \mathrm{argmin}_f \, \mathrm{E}\left(\phi(Yf(\mathbf{X}))\right).$$

4. Consider the M-step in the EM algorithm for the Gaussian mixture model, which minimizes
   $$\underset{\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K}{\mathrm{argmax}} \sum_{i=1}^n \sum_{k=1}^K [\log \phi(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \log \pi_k] \tilde{p}_{ik}^{(t)}.$$
   at the t^{th} step in which $\sum_{k=1}^K \tilde{p}_{ik}^{(t)} = 1$ for each given i. Show that the solution is given by
   $$\begin{aligned} \pi_k^{(t+1)} &= \frac{1}{n} \sum_{i=1}^n \tilde{p}_{ik}^{(t)}, \\ \boldsymbol{\mu}_k^{(t+1)} &= \frac{\sum_{i=1}^n \tilde{p}_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^n \tilde{p}_{ik}^{(t)}}, \\ \boldsymbol{\Sigma}_k^{(t+1)} &= \frac{\sum_{i=1}^n \tilde{p}_{ik}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{i=1}^n \tilde{p}_{ik}^{(t)}}. \end{aligned}$$
   Note that the third step involves complicated calculation of the derivatives of the determinant and inverse of the covariance matrix, which is beyond our class. Therefore, you are not required to prove third identity.

5. The Email Spam dataset consists of 4601 email messages of which about 40% are spam emails. There are also 57 predictors such as frequencies of certain words, total number of capital letters, etc. This dataset is publicly available at the UCI Machine Learning Repository (Bache and Lichman, 2013): `https://archive.ics.uci.edu/ml/datasets/Spambase`.

   The goal is to identify spams from real emails. Let us randomly split the dataset into a training set of 1000 observations and a testing set of 3601 observations (set the random seed to 525 `set.seed(525)` before splitting so that everyone in the class has the same training and testing data). To compute the augmented features, use the R command `density` and its default bandwidth. Report the misclassfication error rates on the testing set by using

   (a) penalized logistic regression;

   (b) penalized logistic regression with augmented features;

(c) linear support vector machine with augmented features;

(d) CART;

(e) random forest.

6. Download the mice protein expression data from UCI Machine Learning Repository:

   `https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression.`

   (a) There are in total 1080 examples. Remove the examples with missing values. Report the total number of examples afterwards. Use attributes 2–78 (the expression levels of 77 proteins) as the input variables for clustering analysis. Attribute 82 shows the true class labels (there are 8 classes).

   (b) Apply k-means clustering, spectral clustering and agglomerative hierarchical clustering to this dataset with the number of clusters $k = 8$. Compare the clustering results to the ground truth. You can use the adjusted Rand index, which measures the fidelity of the clustering with respect to the true labels.

   Note: Technically, we have not learned PCA yet, but the function `specc` in the package 'kernlab' does it for all.