Assignment 5 (3 pages)

Statistics 32950-24620 (Spring 2024)

Due 9 am Tuesday, April 30.

Requirements: Same as before. Make sure to submit to the correct section 246Pset5 or 329Pset5 in Gradescope.

References:

Chapter 11 in Johnson & Wichern. Chapter 2 and sections 4.1-4.3, 12.1-12.2 in Hastie, Tibshirani and Friedman.

Problem assignments:

1. (*2-class mini k-NN clssification*)

   The training data (x_i, y_i) $(i = 1, 2, 3)$ are $(0.3, 1), (0.5, 1), (0.7, 0)$.

   (a) For all x in $[0, 1]$, determine the output of the binary class label y given by a k-Nearest Neighbor (k-NN) classifier, using

   i. 1-NN

   ii. 3-NN

   (b) Using the mean of the k-nearest neighbors of a test point, plot the output y (no longer binary) for all x in $[0, 1]$, by using 2-NN.

   Note: For special or ambivalent assignments, state your reasoning.

2. (*Classification of two overlapping populations*)

   Let $f_1(x) = c_1(1 - |x - 0.5|)\mathbf{1}_{\{-0.5 \le x \le 1.5\}}$, $f_2(x) = c_2(1 - |x|)\mathbf{1}_{\{-1 \le x \le 1\}}$, $f_3(x) = c_3(2 - |x - 0.5|)\mathbf{1}_{\{-1.5 \le x \le 2.5\}}$.

   (a) For $i = 1, 2, 3$, find the values of c_i so that each $f_i(x)$ is a probability density function.

   (b) Identify the classification regions for the two populations with density functions f_1 and f_2 by the classification rule of minimum expected cost of misclassification (ECM), when the prior probability of Population 1 is $p_1 = 0.8$ and the cost of misclassification $c(1|2) = c(2|1)$.
   (Note: It helps to sketch f_1 and f_2, densities for Populations 1 and 2.)

   (c) Sketch the two densities f_1 and f_3, densities for Population 1 and 3 on the same plot. Identify the classification regions by the minimum ECM rule when $p_1 = 0.8$ and $c(1|3) = c(3|1)$.

3. (*Fisher's linear discriminants for three classes*) (Partially based on Exercise 11.29 in J&W.)

   A business school admissions committee used GPA and GMAT scores to make admission decisions.

   The dataset is GpaGmat.DAT (text data, automatic download when clicked, also available next to the link of this p-set in Canvas). The variable `admit = 1,2,3` corresponds to admission decision "Yes", "No", "Borderline".

   ```
   The following R commands can be used to read in the data.

   > gsbdata = read.table("GpaGmat.DAT")
   > colnames(gsbdata)=c("GPA", "GMAT","admit")
   ```

   (a) Calculate $\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{x}}$ and $\boldsymbol{S}_{pool}$. (Note: $\bar{\boldsymbol{x}}$ is the overall average of all obs. Is it the same as the average of the subgroup means?)

   (b) Calculate the sample within-group sum of squares and cross products matrix $\boldsymbol{W}$, its inverse $\boldsymbol{W}^{-1}$, and the sample between-group sum of squares and cross products matrix $\boldsymbol{B}$. (A^{-1} is `solve(A)` in R.)
   (Caution: Be sure to use the right average in $\boldsymbol{B}$.)
   Find the eigenvalues λ_1, λ_2 (or rather the estimates $\hat{\lambda}_1, \hat{\lambda}_2$) and eigenvectors $\boldsymbol{a}_1, \boldsymbol{a}_2$ of $\boldsymbol{W}^{-1}\boldsymbol{B}$.

(c) Use the linear discriminants derived from these eigenvectors to classify two new observations $\boldsymbol{x} = [3.21\ 497]'$ and $\boldsymbol{x} = [3.22\ 497]'$.

Note: You may use the common rule that $\boldsymbol{x}$ is assigned to class π_k if $\boldsymbol{a}'\boldsymbol{x}$ (in $\mathbb{R}^2$) is closest to $\boldsymbol{a}'\bar{\boldsymbol{x}}_k$ for $k = 1, \cdots, g$ (here g=3), where $\boldsymbol{a} = [\boldsymbol{a}_1\ \boldsymbol{a}_2]$ is the 2×2 eigenvector matrix.

(d) Plot the original dataset on the plane of the first two discriminants, labeled by admission decisions. Comment on the results in (c). Is the admission policy a good one?

4. (*Hands-on classifications for two normal populations*)

Two data sets $\boldsymbol{X}_1, \boldsymbol{X}_2$ are sampled from two bivariate normal populations π_1, π_2, with $\boldsymbol{X}_1 = \{[3\ 7]', [2\ 4]', [4\ 7]')\}$, $\boldsymbol{X}_2 = \{[6\ 9]', [5\ 7]', [4\ 8]'\}$.
The data and basic statistics can be obtained by the following R commands.

```
> X1 = cbind(c(3,2,4),c(7,4,7))
> X2=cbind(c(6,5,4),c(9,7,8))
> mu1 = colMeans(X1)
> mu2 = colMeans(X2)
> S1=cov(X1)
> S2=cov(X2)
```

(a) Assuming equal covariance matrices $\Sigma_1 = \Sigma_2$ in the two populations, calculate the estimated linear discriminant function $y = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_{pool}^{-1}\boldsymbol{x}$.

(b) Use the rule of minimum expected costs of misclassification (ECM) to classify the new observations $\boldsymbol{x} = [4.1\ 5]'$ and $\boldsymbol{x} = [3.9\ 9]'$, under the assumption of equal costs and equal priors.
(Under these assumptions the classifier is equivalent to Fisher's linear discriminant function.)

(c) Repeat Part (b) with cost $c(2|1) = \$3, c(1|2) = \$20$, and assuming that about 10% of all possible observations belong to population π_1.

(d) Assuming that $\Sigma_1 \neq \Sigma_2$ in the two bivariate normal distributions. Use the general quadratic rule

$$R_1 = \left\{ \boldsymbol{x} : d(\boldsymbol{x}) \geq \ln\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right) \right\}, \qquad R_2 = \left\{ \boldsymbol{x} : d(\boldsymbol{x}) < \ln\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right) \right\}$$

to classify the observations $\boldsymbol{x}' = [4.1\ 5]'$ and $\boldsymbol{x}' = [3.9\ 9]'$, where

$$d(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}'(\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1})\boldsymbol{x} + (\bar{\boldsymbol{x}}_1'\boldsymbol{S}_1^{-1} - \bar{\boldsymbol{x}}_2'\boldsymbol{S}_2^{-1})\boldsymbol{x} - \hat{k}, \quad \hat{k} = \frac{1}{2}(\bar{\boldsymbol{x}}_1'\boldsymbol{S}_1^{-1}\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2'\boldsymbol{S}_2^{-1}\bar{\boldsymbol{x}}_2) + \frac{1}{2}\ln\frac{|\mathbf{S_1}|}{|\boldsymbol{S}_2|}.$$

Let's make it simple by using equal costs and equal priors as in (b).

(e) Compare and comment on the classification results in (b) and (d).

(f) Test for the difference in population mean vectors using Hotelling's two-sample T^2 test statistic.

5. (*Graphical exercise on linear discriminants and conceptual SVM*)

This geometric-graphical exercise is based on the data in Question 4 in this assignment.

The two data sets sampled from two classes are $\boldsymbol{X}_1 = \{[3\ 7]', [2\ 4]', [4\ 7]')\}, \boldsymbol{X}_2 = \{[6\ 9]', [5\ 7]', [4\ 8]'\}$.

Plot the data points (as (x_{ci}, y_{ci}), $c = 1, 2; i = 1, 2, 3$) on the x-y plane, with clear labels on the plot indicating the class of each point. This will be the base plot for the following parts.

(a) On your base plot, based on your results in parts (a) and (b) in Question 5, plot the linear classification border $\hat{y} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_{pool}^{-1}\boldsymbol{x} - \hat{m} = 0$ obtained by the linear discriminant function, and plot the two new observations $(4.1, 5)$ and $(3.9, 9)$. (Assuming equal covariance $\Sigma_1 = \Sigma_2$, equal costs and equal priors.)

(b) (Conceptual plots only. No calculations or usage of software which would use different, more sophisticated criteria.)
Start with another base plot containing points from samples $\boldsymbol{X}_1, \boldsymbol{X}_2$,

i. Add the linear classifier obtained by the method of SVM.

ii. Identify the supporting vectors.

iii. Plot the two observations $(4.1, 5)$ and $(3.9, 9)$. To which classes are they assigned by linear SVM?

(c) Compare (a) and (b). Which classifier do you prefer? Your reasons?

(d) (**Required for 32950 only**. Optional/bonus for 24620)

In part (d) of Question 4, under the assumptions of equal costs, equal priors, $\Sigma_1 \neq \Sigma_2$, and bivariate normal distributions for the two classes, two new observations $(4.1, 5)$ and $(3.9, 9)$ were classified by the quadratic function $d(\boldsymbol{x})$.

Now for this visualization exercise,

i. On another base plot, plot $d(\boldsymbol{x}) = d(x, y) = 0$, the class border(s) by quadratic discriminant function.
(Hint: It is numerically easier to use $3d(x, y) = 0$, which is a degenerate conic section equation, now with integer coefficients.)

ii. Plot the two observations $(4.1, 5)$ and $(3.9, 9)$.
Which classes are they assigned into by the quadratic discriminant rule?

iii. Classify another new observation $(4.1, 9.5)$. Is the classification reasonable?

iv. Usually we go for higher order (such as from linear to quadratic) and less assumptions for more refined or "better" classification rules. Is the quadratic rule here better than the linear discriminant classifier in part (a)? Can you explain the reason by the pattern of the quadratic classification regions?