IEOR 290 002 - Spring 2017
Homework 6
Due 04/03/2017 at 11:59 pm

Instruction

- We suggest you type your answer using LaTeX or Microsoft Word.
- This homework contains both theoretical and coding parts. For questions with `[coding]` mark, you need to write a small program to answer the question. We recommend you use Python, R or MATLAB to do the problem. Append your code after the answer.
- Submit a single PDF of your write-up on bCourses. Make sure your answer is legible.

1. **(Fisher's linear discriminant)**

   A fisher wants to design an automated fishing system that captures fish, classifies them, and sends them off to two different companies, Salmonites Inc., and Seabass Inc. Suppose the fisher only catches salmon ($t = 1$) and seabass ($t = 2$) and of course Salmonites wants only salmon and SeaBass wants only seabass. Given only the weights of the fish, the fisher wants to figure out what type of fish it is using Fisher's linear discriminant.

   Assume the weight of both salmon and seabass are independently normally distributed as $\mathcal{N}(\mu_t, \sigma_t), t = 1, 2$ and the fisher has the following training data:

   - Weight of salmon: $\{3, 4, 5, 6, 7\}$
   - Weight of seabass: $\{5, 6, 7-\sqrt{2}, 7, 7+\sqrt{2}, 8, 9, \}$

   (a) Compute the sample mean $\hat{\mu}_t$ and sample variance $\hat{\sigma}_t^2$ of both salmon and seabass. (Use $\sigma_t^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_t)^2$ when computing the sample variance)

   (b) Compute the estimates of class priors $\hat{p}_1 = \mathbf{P}(t = 1)$ and $\hat{p}_2 = \mathbf{P}(t = 2)$

   (c) Given a new fish with weight x', define the risk of making a prediction of $\hat{t} = k$ as

   $$R(\hat{t} = k|x') = \sum_{j=1}^{2} l_{jk}\mathbf{P}(t = j|x')$$

   where l_{jk} is the loss of predicting a fish from class j as class k. The optimal decision rule is to predict $\hat{t} = \arg\min_{k=1,2} R(\hat{t} = k|x')$.

Assume the loss is 0/1, i.e.

$$l_{jk} = \begin{cases} 1 & k \neq j \\ 0 & k = j \end{cases}$$

then the optimal decision rule reduces to minimizing the probability of making error. Find the decision boundary x^* such that

$$R(\hat{t} = 1|x^*) = R(\hat{t} = 2|x^*)$$

(d) When the fisher classifies seabass incorrectly, it gets sent to Salmonites who won't pay for the wrong fish and sells it themselves. When the fisher classifies salmon incorrectly, it gets sent to SeaBass who is nice and returns the fish. This situation gives the following loss matrix

l_{jk}		prediction k	
		salmon	seabass
truth j	salmon	0	1
	seabass	2	0

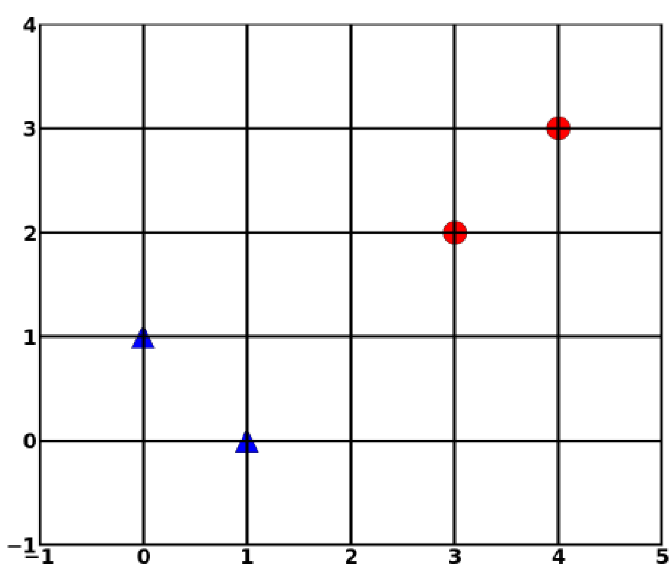
Find the decision boundary x^* such that

$$R(\hat{t} = 1|x^*) = R(\hat{t} = 2|x^*)$$

How the decision boundary x^* changes compared to part (c)?

2. **(Support vector machine)**

Consider the following set of data (blue triangle $= +1$, red dot $= -1$)



Find the equation (by hand) of the hyperplane $\boldsymbol{w}^\top \boldsymbol{x} + w_0$ that would be used by the support vector machine. Which points are support vectors?

3. **(Comparison between Logistic regression, LDA, QDA and SVM)**

   `[Coding]` Let data $\boldsymbol{x} \in \mathbb{R}^2$ be generated from two classes $t = 1$ and $t = -1$ such that

   - $X|t = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ with

   $$\boldsymbol{\mu}_1 = \begin{bmatrix} -2 \\ -2 \end{bmatrix} \quad \boldsymbol{\Sigma_1} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

   - $X|t = -1 \sim \mathcal{N}(\boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}_{-1})$ with

   $$\boldsymbol{\mu}_{-1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \boldsymbol{\Sigma_{-1}} = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}$$

   Generate 200 points from class $t = 1$ and 100 points from class $t = -1$.

   (a) Train a classifier using
   - Logistic regression with l_2 regularization using $\lambda = 0.1$
   - Fisher's linear discriminant
   - Quadratic discriminant analysis
   - Support vector machine with l_2 regularization using $\lambda = 1$ and linear kernel
   - Support vector machine with l_2 regularization using $\lambda = 1$ and Gaussian kernel

   (b) Report the training accuracy

   (c) Plot the decision boundary of the each classifier

   Useful functions in Python

```
sklearn.linear_model.LogisticRegression()
sklearn.discriminant_analysis.LinearDiscriminantAnalysis()
sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis()
sklearn.svm.SVC()
```