INFO-F-422 : Statistical foundations of machine learning
Theory exam : 1st session

Gianluca Bontempi,
Computer Science Department, ULB

This assessment counts for 40% of your grade. Student cannot use the handbook. Only material for writing (e.g. graph paper) and a pocket calculator are accepted.

1 Question (2 points)

1. Mention in which case a combination of estimators can be beneficial
2. Prove it analytically *(Use a red ink to identify the random variables in the demonstration).*
3. Discuss at least 2 techniques that are based on this principle (at most 10 lines per technique : use of pseudo code is welcome)
   Solution : The student could discuss for instance the bagging and the Random Forest learning technique.

2 Question (1 point)

Derive the bias/variance/noise decomposition of the test error in a regression problem. Use a red ink to identify the random variables in the demonstration.

3 Question (1,5 points)

Let us consider a fraud detection problem. Suppose we collect the following transactional dataset where $\mathbf{v} = 1$ means that the transaction came from a suspicious web site and $\mathbf{f} = 1$ means that the transaction is fraudulent.

	$\mathbf{f} = 1$	$\mathbf{f} = 0$
$\mathbf{v} = 1$	500	1000
$\mathbf{v} = 0$	1	10000

1. Estimate the following quantities by using the frequency as estimator of probability :

— Prob $\{\mathbf{f}=1\}$
Solution : Prob $\{\mathbf{f}=1\} = 501/11501 = 0.043$

— Prob $\{\mathbf{v}=0\}$
Solution : Prob $\{\mathbf{v}=0\} = 10001/11501 = 0.869$

— Prob $\{\mathbf{f}=1|\mathbf{v}=1\}$
Solution : Prob $\{\mathbf{f}=1|\mathbf{v}=1\} = 500/1500 = 1/3$

— Prob $\{\mathbf{v}=1|\mathbf{f}=1\}$
Solution : Prob $\{\mathbf{f}=1|\mathbf{v}=1\} = 500/501$

2. Use the Bayes theorem to compute Prob $\{\mathbf{v}=1|\mathbf{f}=1\}$ and show that the result is identical to the one computed before.
Solution : Prob $\{\mathbf{f}=1|\mathbf{v}=1\} = \frac{1/3(1500/11501)}{501/11501} = 500/501$
3. Suppose you make a classifier that returns always $\mathbf{f}=1$: what would have been its specificity, sensitivity and precision for the collected dataset ?
Solution :

	P	N
$\hat{P}$	TP=501	FP=11000
$\hat{N}$	FN=0	TN=0

SE=TP/(TP+FN)=1
SP=TN/(FP+TN)=0
PR=TP/(TP+FP)=501/11501

4 Question (2 points)

Consider a regression task with two inputs $\mathbf{x}_1$, $\mathbf{x}_2$ and output $\mathbf{y}$. Suppose we observe the following training set

X_1	X_2	Y
-0.2	0 .1	1
0.1	0	0.5
1	-0.3	1.2
0.1	0.2	1
-0.4	0.4	0.5
0.1	0.1	0
1	-1	1.1

1. Fit a multivariate linear model with $\beta_0 = 0$ to the dataset.
Solution : $X^TX = \begin{bmatrix} 2.23 & -1.45 \\ -1.45 & 1.31 \end{bmatrix}$

$(X^TX)^{-1} = \begin{bmatrix} 1.599 & 1.77 \\ 1.77 & 2.72 \end{bmatrix}$

$X^TY = \begin{bmatrix} 2.05 \\ -0.96 \end{bmatrix}$

$\beta = (X^TX)^{-1}X^TY = \begin{bmatrix} 1.58 \\ 1.016 \end{bmatrix}$

2. Compute the mean squared training error.

Solution : $e = Y - X\beta = \begin{bmatrix} 1.21 \\ 0.34 \\ -0.08 \\ 0.64 \\ 0.73 \\ -0.26 \\ 0.54 \end{bmatrix}$

MSE=0.41

3. Suppose you use a correlation-based ranking strategy for ranking the features. What would be the top ranked variable?

Solution : Since

$$\rho_{X_1Y} = \frac{\sum_{i=1}^N (X_{i1} - \mu_1)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^N (X_{i1} - \mu_1)^2 (Y_i - \mu_Y)^2}} = 0.53$$

and

$$\rho_{X_2Y} = \frac{\sum_{i=1}^N (X_{i2} - \mu_2)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^N (X_{i2} - \mu_2)^2 (Y_i - \mu_Y)^2}} = -0.48$$

where $\mu_1 = 0.24, \mu_2 = -0.07, \mu_Y = 0.75$, X_1 is the top ranked variable.

Hint :

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \Rightarrow A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}^2} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{12} & a_{11} \end{bmatrix}$$

5 Question (3,5 points)

Let us consider a classification task with 3 binary inputs and one binary output. Suppose we collected the following training set

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{y}$
1	1	0	1
0	0	1	0
0	1	0	0
1	1	1	1
0	0	0	0
0	1	0	0
0	1	1	0
0	0	1	0
0	0	0	0
0	1	0	0
1	1	1	1

1. Estimate the following quantities by using the frequency as estimator of probability

— Prob $\{\mathbf{y}=1\}$
Solution : Prob $\{\mathbf{y}=1\}=3/11$
— Prob $\{\mathbf{y}=1|\mathbf{x}_1=0\}$
Solution : Prob $\{\mathbf{y}=1|\mathbf{x}_1=0\}=0$
— Prob $\{\mathbf{y}=1|\mathbf{x}_1=0,\mathbf{x}_2=0,\mathbf{x}_3=0\}$
Solution : Prob $\{y=1|\mathbf{x}_1=0,\mathbf{x}_2=0,\mathbf{x}_3=0\}=0$

2. Consider a Naive Bayes classifier and compute its classifications if the same dataset is used also for testing
Solution : Note that the values of $\mathbf{x}_1$ are identical to the ones of $\mathbf{y}$. Then Prob $\{\mathbf{x}_1=A|\mathbf{y}=\neg A\}=0$. It follows that if use a Naive Bayes and the test dataset is equal to the training set all the predictions will coincide with the values of $\mathbf{x}_1$. The training error is then zero
3. Trace the ROC curve associated to the Naive Bayes classifier if the same dataset is used also for testing. (Hint : make the assumption that the denominator of the Bayes formula is 1 for all test points)
Solution : Since all the predictions are correct the ROC curve is equal to 1 for all FPR values

Bonus question (1,5 points)

The points of the bonus question will be taken into account only if the students got at least 8 points in the previous questions

Consider the data set in Question 4 and let us fit to it a radial basis function with 2 basis functions having as parameters $\mu^{(1)}=[0,0]$ and $\mu^{(2)}=[1,1]$. The equation of the basis function is

$$\rho(x,\mu)=\prod_{i=1}^{2} exp^{-(x_i-\mu_i)^2}$$

where x_i (μ_i) stands for the ith coordinate of the vector x (μ).

The student should return the equation of the radial basis function.

Solution : The equation of the RBF is $h(x)=\sum_{j=1}^{2} w_j\rho^{(j)}(x,\mu^{(j)})$

The only unknown terms are the weights w_1 and w_2 which should be estimated by least-squares solving the equation $Y=RW$ where $W=\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ and

$$R=\begin{bmatrix} \rho^{(1)}(x_1,\mu^{(1)}) & \rho^{(2)}(x_1,\mu^{(2)}) \\ \dots & \\ \rho^{(1)}(x_N,\mu^{(1)}) & \rho^{(2)}(x_N,\mu^{(2)}) \end{bmatrix}$$

For our dataset

$$R = \begin{bmatrix} 0.95 & 0.11 \\ 0.99 & 0.16 \\ 0.34 & 0.18 \\ 0.95 & 0.23 \\ 0.73 & 0.10 \\ 0.98 & 0.20 \\ 0.14 & 0.02 \end{bmatrix}$$

By solving the least-squares $W = (R^T R)^{-1} R^T Y$ we obtain $w_1 = 0.02, w_2 = 3.94$