INFO-F-422 : Statistical foundations of machine learning
Questions on theory.

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 (1 point)

Define first the bias and variance of an estimator $\hat{\boldsymbol{\theta}}$. Then demonstrate the bias/variance decomposition of the Mean Squared Error. Use a red ink to identify the random variables in the demonstration.

1.1 Solution

See slide 27 of *Parametric approaches to estimation*

2 Question (1 point)

Mention in which case a combination of estimators can be beneficial and prove it analytically. Use a red ink to identify the random variables in the demonstration.

2.1 Solution

See slides 22-23 of *Non parametric approaches to estimation*

3 Question (2 points)

Derive analytically the bias and the variance of the sample average estimator. Use a red ink to identify the random variables in the demonstration.

3.1 Solution

See slide 21 of *Parametric approaches to estimation*

4 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.

4.1 Solution

See slides 11-12 of *Nonlinear models (Supervised learning)*

5 Question (2 point)

Write down the pseudo-code of a structural identification procedure based on leave-one-out.

5.1 Solution

See slides 28 of *Nonlinear models (Supervised learning)*

6 Question (3 points)

Let us consider the following observations of the random variable $\mathbf{z}$

$$D_N = \{0.1, -1, 0.3, 1.4\}$$

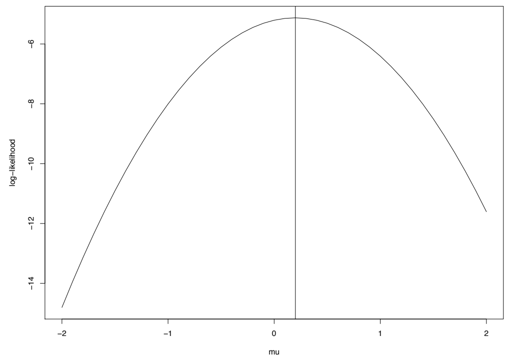
Write the analytical form of the likelihood function of the mean μ for a Gaussian distribution with a variance $\sigma^2 = 1$. The student should :

1. Trace the log-likelihood function on the graph paper
2. Determine graphically the maximum likelihood estimator.
3. Discuss the result.

6.1 Solution

Since $N = 4$ and $\sigma = 1$

$$L(\mu) = \prod_{i=1}^{N} p(\mathbf{z}_i, \mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp \frac{-(z_i - \mu)^2}{2}$$



Note that $\hat{\mu}_{ml}$ coincides with the sample average $\hat{\mu} = 0.2$ of D_N.

7 Question (2 points)

Let us consider a spam detection problem. Suppose we collect the following data about the received emails where $\mathbf{v} = 1$ stands for the presence of the word Viagra in the email text and $\mathbf{s} = 1$ stands for the user classification of the email as spam.

	$\mathbf{s} = 1$	$\mathbf{s} = 0$
$\mathbf{v} = 1$	20	10
$\mathbf{v} = 0$	5	10

1. Estimate the following quantities by using the frequency as estimator of probability :
   — Prob $\{\mathbf{s} = 1\}$
   — Prob $\{\mathbf{v} = 0\}$
   — Prob $\{\mathbf{s} = 1 | \mathbf{v} = 1\}$
   — Prob $\{\mathbf{v} = 1 | \mathbf{s} = 1\}$
2. Use the Bayes theorem to compute Prob $\{\mathbf{v} = 1 | \mathbf{s} = 1\}$ and show that the result is identical to the one computed before.
3. Suppose you receive an email containing the word Viagra : how would you classify this email (spam or no spam) supposing that the cost of a false positive is equal to the cost of a false negative ?

7.1 Solution

1. — $\widehat{\text{Prob}}\{\mathbf{s} = 1\} = \frac{25}{45} = \frac{5}{9}$
   — $\widehat{\text{Prob}}\{\mathbf{v} = 0\} = \frac{15}{45} = \frac{1}{3}$
   — $\widehat{\text{Prob}}\{\mathbf{s} = 1 | \mathbf{v} = 1\} = \frac{20}{30} = \frac{2}{3}$
   — $\widehat{\text{Prob}}\{\mathbf{v} = 1 | \mathbf{s} = 1\} = \frac{20}{25} = \frac{4}{5}$

2. $\widehat{\text{Prob}}\{\mathbf{v}=1|\mathbf{s}=1\} = \frac{\widehat{\text{Prob}}\{\mathbf{s}=1|\mathbf{v}=1\}\widehat{\text{Prob}}\{\mathbf{v}=1\}}{\widehat{\text{Prob}}\{\mathbf{s}=1\}} = \frac{\frac{2}{3}\frac{2}{3}}{\frac{5}{9}} = \frac{4}{5}$

3. $\widehat{\text{Prob}}\{\mathbf{s}=1|\mathbf{v}=1\} = \frac{2}{3} > 0.5$ then the email should be classified as a spam.

8 Question (2 points)

Let us consider the following joint probability of three random binary variables

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	P
0	0	0	0.2
0	0	1	0.1
0	1	0	0.05
0	1	1	0.1
1	0	0	0.05
1	0	1	0.1
1	1	0	0.05
1	1	1	0.25

1. Compute
   — $\text{Prob}\{\mathbf{x}_1=1, \mathbf{x}_2=1\}$
   — $\text{Prob}\{\mathbf{x}_1=0|\mathbf{x}_2=1, \mathbf{x}_3=0\}$
   — $\text{Prob}\{\mathbf{x}_1=0|\mathbf{x}_2=1\}$
   — $\text{Prob}\{\mathbf{x}_3=1|\mathbf{x}_2=1\}$
2. Use the Bayes theorem to compute $\text{Prob}\{\mathbf{x}_1=0|\mathbf{x}_2=1, \mathbf{x}_3=0\}$ and compare the result to the one computed before.

8.1 Solution

1. — $\text{Prob}\{\mathbf{x}_1=1, \mathbf{x}_2=1\} = 0.3$
   — $\text{Prob}\{\mathbf{x}_1=0|\mathbf{x}_2=1, \mathbf{x}_3=0\} = \frac{0.05}{0.1} = 1/2$
   — $\text{Prob}\{\mathbf{x}_1=0|\mathbf{x}_2=1\} = \frac{0.15}{0.45} = 1/3$
   — $\text{Prob}\{\mathbf{x}_3=1|\mathbf{x}_2=1\} = \frac{0.35}{0.45} = 7/9$
2. $\text{Prob}\{\mathbf{x}_1=0|\mathbf{x}_2=1, \mathbf{x}_3=0\} = \frac{\text{Prob}\{\mathbf{x}_2=1,\mathbf{x}_3=0|\mathbf{x}_1=0\}\text{Prob}\{\mathbf{x}_1=0\}}{\text{Prob}\{\mathbf{x}_2=1,\mathbf{x}_3=0\}} = \frac{\frac{0.05}{0.45}0.45}{0.1} = 1/2$

9 Question (3 points)

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

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

1. Estimate the following quantities by using the frequency as estimator of probability
   — $\text{Prob}\{\mathbf{y}=1\}$
   — $\text{Prob}\{\mathbf{y}=1|\mathbf{x}_1=0\}$
   — $\text{Prob}\{\mathbf{y}=1|\mathbf{x}_1=0,\mathbf{x}_2=0,\mathbf{x}_3=0\}$
2. Compute the classification returned by using the Naive Bayes Classifier for the input $\mathbf{x}_1=0,\mathbf{x}_2=0,\mathbf{x}_3=0$.
3. Suppose we test a classifier for this task and that we obtain a misclassification error equal to 20%. Is it working better than a zero classifier, i.e. a classifier ignoring the value of the inputs ?

9.1 Solution

Let us note that $N=12$

1. — $\widehat{\text{Prob}}\{\mathbf{y}=1\}=2/12=1/6$
   — $\widehat{\text{Prob}}\{\mathbf{y}=1|\mathbf{x}_1=0\}=\frac{1}{6}$
   — $\widehat{\text{Prob}}\{\mathbf{y}=1|\mathbf{x}_1=0,\mathbf{x}_2=0,\mathbf{x}_3=0\}$ cannot be estimated using the frequency since there is no observation where $\mathbf{x}_1=0,\mathbf{x}_2=0,\mathbf{x}_3=0$
2. Since

$$\widehat{\text{Prob}}\{\mathbf{y}=1|\mathbf{x}_1=0,\mathbf{x}_2=0,\mathbf{x}_3=0\}\propto$$
$$\widehat{\text{Prob}}\{\mathbf{x}_1=0|\mathbf{y}=1\}\widehat{\text{Prob}}\{\mathbf{x}_2=0|\mathbf{y}=1\}\widehat{\text{Prob}}\{\mathbf{x}_3=0|\mathbf{y}=1\}\widehat{\text{Prob}}\{\mathbf{y}=1\}=$$
$$(0.5*0.5*0.5*1/6)=0.02$$

and

$$\widehat{\text{Prob}}\{\mathbf{y}=0|\mathbf{x}_1=0,\mathbf{x}_2=0,\mathbf{x}_3=0\}\propto$$
$$\widehat{\text{Prob}}\{\mathbf{x}_1=0|\mathbf{y}=0\}\widehat{\text{Prob}}\{\mathbf{x}_2=0|\mathbf{y}=0\}\widehat{\text{Prob}}\{\mathbf{x}_3=0|\mathbf{y}=0\}\widehat{\text{Prob}}\{\mathbf{y}=0\}=$$
$$(5/10*4/10*5/10*5/6)=0.08$$

the NB classification is 0

3. A zero classifier would return always the class with the highest a priori probability, that is the class 0. Its misclassification error would be then $1/6$. Since $1/5 > 1/6$ the classifier is working worse than the zero classifier.

10 Question (2 points)

Consider a regression task with input $\mathbf{x}$ and output $\mathbf{y}$. Suppose we observe the following training set

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

and that the prediction model is constant. Compute an estimation of its mean integrated squared error by leave-one-out.

10.1 Solution

Since the leave-one-out error is

$$e_i^{-i} = y_i - \frac{\sum_{j=1,j\neq i}^{N} y_j}{N-1}$$

we can compute the vector of errors in leave-one-out

e_1^{-1}	1- 0.716=0.283
e_2^{-2}	0.5- 0.8= -0.3
e_3^{-3}	1.2- 0.683=0.516
e_4^{-4}	1- 0.716=0.283
e_5^{-5}	0.5- 0.8= -0.3
e_6^{-6}	0- 0.883=-0.883
e_7^{-7}	1.1- 0.7=0.4

and then derive the MISE estimation

$$\widehat{\text{MISE}}_{\text{loo}} = \frac{\sum_{i=1}^{N}(e_i^{-i})^2}{N} = 0.22$$

11 Question (3 points)

Consider a regression task with input $\mathbf{x}$ and output $\mathbf{y}$. Suppose we observe the following training set

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

1. Fit a linear model to the dataset.
2. Trace the data and the linear regression function on graph paper.
3. Are the two variables positively or negatively correlated ?

Hint :

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

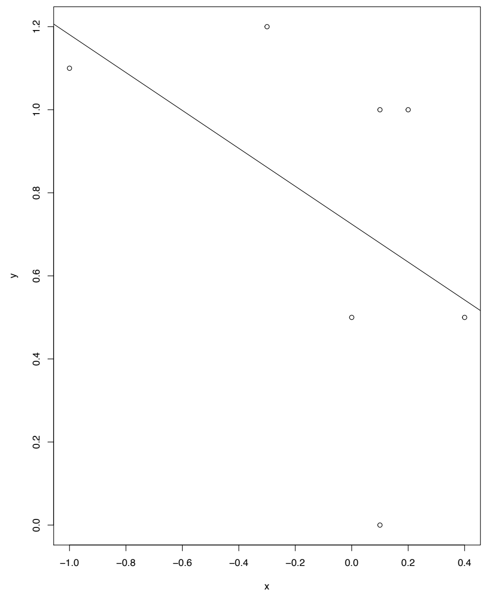
11.1 Solution

1. Once we set
   $$X = \begin{bmatrix} 1 & 0.1 \\ 1 & 0 \\ 1 & -0.3 \\ 1 & 0.2 \\ 1 & 0.4 \\ 1 & 0.1 \\ 1 & -1 \end{bmatrix} \text{ we have}$$

   $$X'X = \begin{bmatrix} 7.0 & -0.50 \\ -0.5 & 1.31 \end{bmatrix}$$

   and

   $$\beta = (X'X)^{-1}X'Y = \begin{bmatrix} 0.725 \\ -0.456 \end{bmatrix}$$



2.

3. Since $\hat{\beta}_1 < 0$ the two variables are negatively correlated.

12 Question (2 points)

Consider a regression task with input $\mathbf{x}$ and output $\mathbf{y}$. Suppose we observe the following training set

X	Y
0 .1	1
0	0.5
-0.3	1.2
0.3	1
0.4	0.5
0.1	0
-1	1.1

and that the prediction model is a KNN (nearest neighbour) where $K = 1$ and the distance metric is euclidean. Compute an estimation of its mean squared error by leave-one-out.

12.1 Solution

The leave-one-out error is

$$e_i^{-i} = y_i - y_i^*$$

where y_i^* is the value of the target associated to x_i^* and x_i^* is the nearest neighbor of x_i. Once we rank the training set according to the input value

X	Y
-1	1.1
-0.3	1.2
0	0.5
0 .1	1
0.1	0
0.3	1
0.4	0.5

we can compute the vector of errors in leave-one-out

e_1^{-1}	1.1-1.2=-0.1
e_2^{-2}	1.2- 0.5= 0.7
e_3^{-3}	0.5- 1=-0.5
e_4^{-4}	1- 0=1
e_5^{-5}	0- 1= -1
e_6^{-6}	1-0.5=0.5
e_7^{-7}	0.5- 1=-0.5

and then derive the MISE estimation

$$\widehat{\text{MISE}}_{\text{loo}} = \frac{\sum_{i=1}^{N}(e_i^{-i})^2}{N} = 0.464$$

13 Question (2 points)

Consider a regression task with input $\mathbf{x}$ and output $\mathbf{y}$. Suppose we observe the following training set

X	Y
0.5	1
1	1
-1	1
-0.25	1
0	0.5
0.1	0
0.25	0.5

Trace the estimation of the regression function returned by a KNN (nearest neighbor) where $K = 3$ on the interval $[-2, 1]$.

13.1 Solution

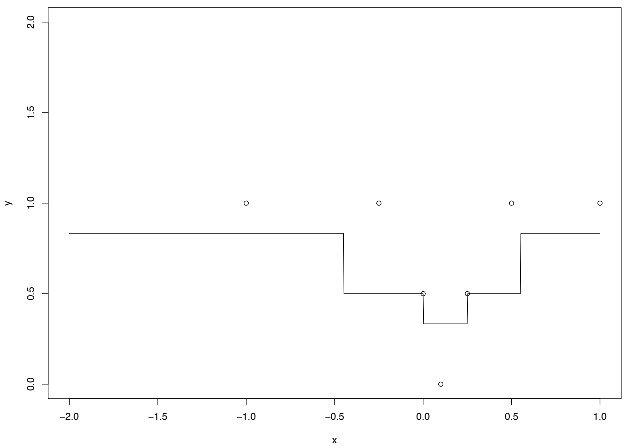
The resulting graph is piecewise constant and each piece has an ordinate equal to the mean of three points. Once ordered the points according to the abscissa

	X	Y
x_1	-1	1
x_2	-0.25	1
x_3	0	0.5
x_4	0.1	0
x_5	0.25	0.5
x_6	0.5	1
x_7	1	1

these are the five sets of 3 points

$$x_1, x_2, x_3 \Rightarrow \hat{y} = 2.5/3 \tag{1}$$
$$x_2, x_3, x_4 \Rightarrow \hat{y} = 0.5 \tag{2}$$
$$x_3, x_4, x_5 \Rightarrow \hat{y} = 1/3 \tag{3}$$
$$x_4, x_5, x_6 \Rightarrow \hat{y} = 0.5 \tag{4}$$
$$x_5, x_6, x_7 \Rightarrow \hat{y} = 2.5/3 \tag{5}$$

The transitions from x_i, x_{i+1}, x_{i+2} to $x_{i+1}, x_{i+2}, x_{i+3}, i = 1, \ldots, 4$ occur at the $x = q$ points where $q - x_i = x_{i+3} - q \Rightarrow q = \frac{x_{i+3}+x_i}{2}$



14 Question (2 points)

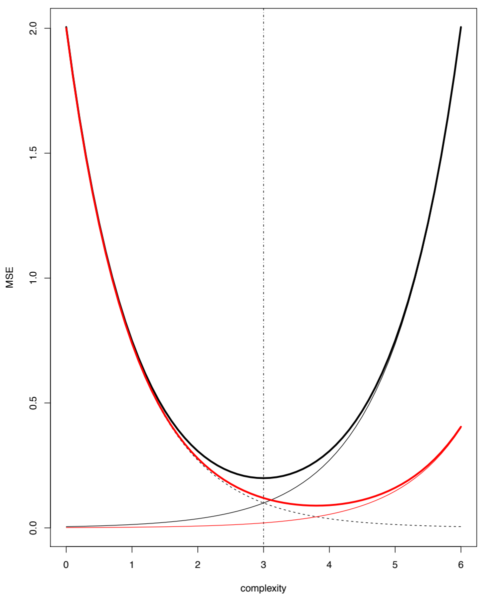
Consider a supervised learning problem, a training set of size $N = 50$ and a neural network predictor with a single hidden layer. Suppose that we are able to compute the generalization error for different number H of hidden nodes and we discover that the lowest generalization error occurs for $H = 3$. Suppose now that the size of the training set increases ($N = 500$). For which value of H would you expect the lowest generalization error ? Equal, larger or smaller than 3 ? Justify your answer by using the bias/variance graphics.

14.1 Solution

MSE=Bias^2+Variance. In the plot we trace the original setting in black (Bias^2 dashed line, Variance continuos line, MSE thick line)

If the training set size increases we can expect a variance reduction. This means that the minimum of the MSE term will move to right. I would expect than a $H > 3$

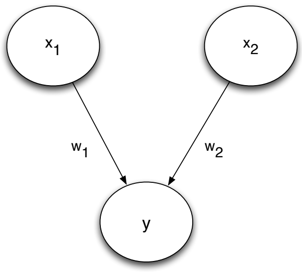
In the plot we trace the new setting in red (Bias^2 unchanged, Variance continuos red thin line, MSE red thick line). It is evident that the $\arg\min MSE$ moved to the right.



15 Question (2 points)

Consider a feedforward neural network with two inputs, no hidden layer and a logistic activation function. Suppose we want to use backpropagation to compute the weights w_1 and w_2 and that a training dataset is collected. The student should

1. Write the equation of the mapping between x_1, x_2 and y.
2. Write the two iterative backpropagation equations to compute w_1 and w_2.



15.1 Solution

1. $\hat{y} = g(z) = g(w_1 x_1 + w_2 x_2)$ where $g(z) = \frac{1}{1+e^{-z}}$ and $g'(z) = \frac{e^{-z}}{(1+e^{-z})^2}$
2. The training error is
$$E = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}$$
For $j = 1, 2$
$$\frac{\partial E}{\partial w_j} = -\frac{2}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial w_j}$$
where
$$\frac{\partial \hat{y}_i}{\partial w_j} = g'(z_i) x_{ij}$$
where $z_i = w_1 x_{1i} + w_2 x_{2i}$
The two backpropagation equations are then
$$w_j(k+1) = w_j(k) + \eta \frac{2}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i) g'(z_i) x_{ij}, \qquad j = 1, 2$$

16 Question (2 points)

Consider a binary classification problem and the following estimations of the conditional probability $\widehat{\text{Prob}}\{\mathbf{y} = 1|x\}$ vs. the real value of the target.

Trace a precision recall and the AUC curve

$\widehat{\text{Prob}}\{\mathbf{y}=1\|x\}$	CLASS
0.6	1
0.5	-1
0.99	1
0.49	-1
0.1	-1
0.26	-1
0.33	1
0.15	-1
0.05	-1

16.1 Solution

Let us first order the dataset in terms of ascending score

$\widehat{\text{Prob}}\{\mathbf{y}=1\|x\}$	CLASS
0.05	-1
0.10	-1
0.15	-1
0.26	-1
0.33	1
0.49	-1
0.50	-1
0.60	1
0.99	1

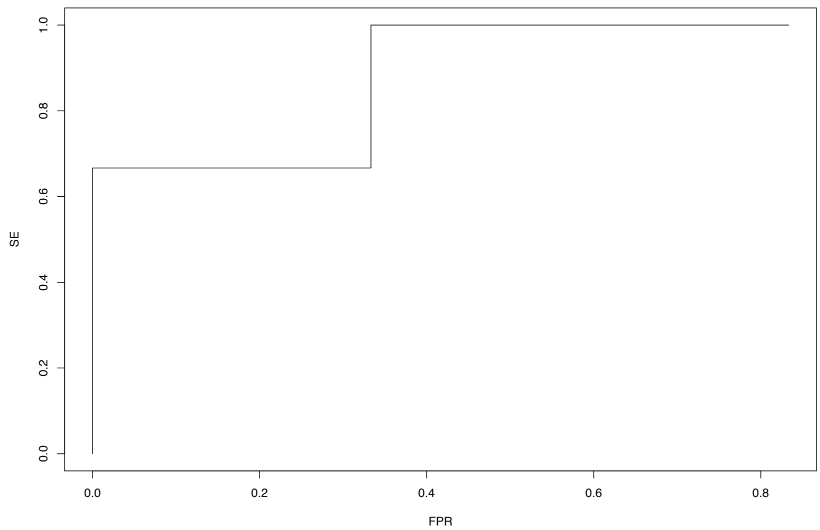
We let the threshold range over all the values of the score. For each value of the threshold we define as positively classified the terms having a score bigger than the threshold and negatively classified the terms having a score lower equal than the threshold.

For instance for Thr=0.26 this is the returned classification

$\widehat{\text{Prob}}\{\mathbf{y}=1\|x\}$	$\hat{y}$	CLASS
0.05	-1	-1
0.10	-1	-1
0.15	-1	-1
0.26	-1	-1
0.33	1	1
0.49	1	-1
0.50	1	-1
0.60	1	1
0.99	1	1

Then we measure the quantity of TP, FP, TN and FN and $FPR = FP/(TN+FP)$, $FPR = TP/(TP+FN)$

Threshold	TP	FP	TN	FN	FPR	TPR
0.05	3	5	1	0	5/6	1
0.10	3	4	2	0	2/3	1
0.15	3	3	3	0	1/2	1
0.26	3	2	4	0	1/3	1
0.33	2	2	4	1	1/3	2/3
0.49	2	1	5	1	1/6	2/3
0.50	2	0	6	1	0	2/3
0.60	1	0	6	2	0	1/3
0.99	0	0	6	3	0	0



17 Question (2 points)

Consider a binary classification problem, a test set and a confusion matrix. Write the formulas of

1. misclassification error
2. balanced error
3. precision
4. recall

as a function of the elements of the confusion matrix (TP,TN,FP,FN).

17.1 Solution

See slides 44-49 of *Classification*

18 Question (2 points)

Consider the dataset

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

Rank the input features in a decreasing order of relevance by using the correlation

$$\rho_{\mathbf{xy}} = \frac{\hat{\sigma}_{\mathbf{xy}}}{\hat{\sigma}_{\mathbf{x}}\hat{\sigma}_{\mathbf{y}}}$$

as measure of relevance.

18.1 Solution

Since $\rho_{\mathbf{x}_1\mathbf{y}} = 0.488$, $\rho_{\mathbf{x}_2\mathbf{y}} = -0.293$, $\rho_{\mathbf{x}_3\mathbf{y}} = -0.192$, the ranking is $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$.