Homework 3 – CS 5525

1. Consider the given joint probability distribution in r.v. X and Y, where

$$\begin{aligned} p(X=0,Y=0) &= \alpha \\ p(X=0,Y=1) &= \beta \\ p(X=1,Y=0) &= \gamma \\ p(X=1,Y=1) &= \delta \end{aligned}$$

(i) Find the expected value of X and Y

Solution : We have

$$p(X=0) = p(X=0,Y=0) + p(X=0,Y=1) = \alpha + \beta$$

Similarly, we have

$$p(X=1) = p(X=1,Y=0) + p(X=1,Y=1) = \gamma + \delta$$

Therefore, the expected value of X is

$$\mathbb{E}(X) = \sum xp(X=x) = 0(\alpha+\beta) + 1(\gamma+\delta) = \gamma + \delta.$$

Calculating similarly for Y, we get

$$\mathbb{E}(Y) = p(Y=1) = \beta + \delta.$$

Remark: There may be multiple ways of expressing the final answer because of the fact that $\alpha+\beta+\gamma+\delta = 1$.

(ii) Find the necessary and sufficient conditions so that X and Y are independent

Solution : From the given joint probability distribution, we can compute the marginals as

$$\begin{aligned} p(X=0) &= \alpha+\beta \\ p(X=1) &= \gamma+\delta \\ p(Y=0) &= \alpha+\gamma \\ p(Y=1) &= \beta+\delta \end{aligned}$$

The necessary and sufficient conditions for independence can be derived using the condition $p(X = x, Y = y) = p(X = x)p(Y = y),\ \forall x, y$. These conditions are listed below

$$\begin{aligned}
p(X = 0, Y = 0) &= \alpha = (\alpha + \beta)(\alpha + \gamma)\\
p(X = 0, Y = 1) &= \beta = (\alpha + \beta)(\beta + \delta)\\
p(X = 1, Y = 0) &= \gamma = (\gamma + \delta)(\alpha + \gamma)\\
p(X = 1, Y = 1) &= \delta = (\gamma + \delta)(\beta + \delta)
\end{aligned}$$

Solving these conditions and using $\alpha+\beta+\gamma+\delta = 1$ simplifies to the condition $\alpha\delta = \beta\gamma$.

——————————————————X——————————————————-

2. **Bayes Theorem & Naïve Bayes Classifier**.
i) Consider a study to determine the effectiveness of a new drug against an infectious disease. There were 10000 test subjects, some of whom were given the real drug while the rest were given a placebo. At the end of the study, 65% of the test subjects recovered from the disease, of out whom half of them took the real drug. Among the test subjects who did not recover from the disease, more than half of them (55%) took the real drug. Based on this information, will taking the drug help a patient to recover from the disease? Also, find the proportion of test subjects who were given the real drug. Show your steps clearly.

Solution : Let $D = 1$ represent population given real drug and $D = 0$ represent the ones given placebo. Also, let $R = 1$ denote recovered and $R = 0$, otherwise. The total number of test subjects is $N = 10000$. We are given the following data

$$\begin{aligned}
&p(R = 1) = 0.65\\
&p(R = 1, D = 1) = 0.5 * 0.65 = 0.325\\
&p(R = 0) = 1 - 0.65 = 0.35\\
&p(R = 0, D = 1) = 0.55 * 0.35 = 0.1925
\end{aligned}$$

Thus, the ratio of the population that were drugged is $p(D = 1) = p(R = 0, D = 1) + p(R = 1, D = 1) = 0.325 + 0.1925 = 0.5175$. Thus, we have that 51.75% of the population was given the real drug. To find the efficacy of drug, we find the conditional probabilities

$$p(R = 1|D = 1) = \frac{p(R = 1, D = 1)}{p(D = 1)} = 0.325/0.5175 \approx 0.6280.$$

and

$$p(R = 1|D = 0) = \frac{p(R = 1, D = 0)}{p(D = 0)} = \frac{0.65 - 0.325}{1 - 0.5175} \approx 0.6736.$$

Thus, the chances of recovery are better if a person does not take the drug.

ii) Consider the given distribution for a binary classification problem using three features.
ii-a) Determine whether X_1 and X_2 are independent of each other.

Solution : We have total 100 samples where $X_1 = 1$ for 65 samples and $X_2 = 1$ for 41 samples. We next verify independence using the marginal distributions and given data as

$$\begin{aligned}
&p(X_1 = 1, X_2 = 1) = 0.28, \quad && p(X_1 = 1)p(X_2 = 1) = 0.65 * 0.41 = 0.2665\\
&p(X_1 = 1, X_2 = 0) = 0.37, &&\\
&p(X_1 = 0, X_2 = 1) = 0.13, &&\\
&p(X_1 = 0, X_2 = 0) = 0.22. &&
\end{aligned}$$

We could have computed the other three products.
However, since $p(X_1 = 1, X_2 = 1) \neq p(X_1 = 1)p(X_2 = 1)$, we know that X_1 and X_2 are not independent.

ii-b) Determine whether X_1 and X_2 are conditionally independent of each other given the class.

Solution : We have 50 positive $P = 1$ and 50 negative $P = 0$ data points. To verify conditional independence, we compute the following probabilities for positive class first

$$\begin{aligned}
&p(X_1 = 1, X_2 = 1|P = 1) = 0.4, \quad && p(X_1 = 1|P = 1)p(X_2 = 1|P = 1) = 0.8 * 0.5 = 0.4\\
&p(X_1 = 1, X_2 = 0|P = 1) = 0.4, && p(X_1 = 1|P = 1)p(X_2 = 0|P = 1) = 0.8 * 0.5 = 0.4\\
&p(X_1 = 0, X_2 = 1|P = 1) = 0.1, && p(X_1 = 0|P = 1)p(X_2 = 1|P = 1) = 0.2 * 0.5 = 0.1\\
&p(X_1 = 0, X_2 = 0|P = 1) = 0.1, && p(X_1 = 0|P = 1)p(X_2 = 0|P = 1) = 0.2 * 0.5 = 0.1
\end{aligned}$$

We can observe that X_1 and X_2 are independent when conditioned on $P = 1$. Let us next verify for the negative class $P = 0$ now

$$\begin{aligned}
&p(X_1 = 1, X_2 = 1|P = 0) = 0.16, \quad && p(X_1 = 1|P = 0)p(X_2 = 1|P = 0) = 0.5 * 0.32 = 0.16\\
&p(X_1 = 1, X_2 = 0|P = 0) = 0.34, && p(X_1 = 1|P = 0)p(X_2 = 0|P = 0) = 0.5 * 0.68 = 0.34\\
&p(X_1 = 0, X_2 = 1|P = 0) = 0.16, && p(X_1 = 0|P = 0)p(X_2 = 1|P = 0) = 0.5 * 0.32 = 0.16\\
&p(X_1 = 0, X_2 = 0|P = 0) = 0.34, && p(X_1 = 0|P = 0)p(X_2 = 0|P = 0) = 0.5 * 0.68 = 0.34
\end{aligned}$$

Thus, from the above two sets of computations, we can conclude that the variables X_1 and X_2 are conditionally independent given the class because $p(X_1, X_2|P) = p(X_1|P)p(X_2|P)$.

ii-c) Compute the class conditional probabilities.

Solution : The needed conditional probabilities are computed below based on instance counting from the provided table

$$\begin{aligned}
p(X_1 = 1| + P = 1) &= 40/50 = 0.8 \\
p(X_1 = 1| - P = 0) &= 25/50 = 0.5 \\
p(X_2 = 1| + P = 1) &= 25/50 = 0.5 \\
p(X_2 = 1| - P = 0) &= 16/50 = 0.32 \\
p(X_3 = 1| + P = 1) &= 20/50 = 0.4 \\
p(X_3 = 1| - P = 0) &= 8/50 = 0.16
\end{aligned}$$

ii-d) Use the class conditional probabilities given in the previous question to predict the class label of each example with the feature set given in the training set above. Use your results to compute the training error of the naïve Bayes classifier.

Solution : From the given data, we have the probabilities of positive class $p(P = 1) = 0.5$ and negative class $p(P = 0) = 0.5$. Since the class probabilities are equal, for Bayesian inference, we can choose a class that maximizes $p(X_1, X_2, X_3|P)$ or assuming conditional independence, $p(X_1|P)p(X_2|P)p(X_3|P)$. The predicted labels of training set are:
Scenario A: $X_1 = 1, X_2 = 1, X_3 = 1$: We have $p(X_1 = 1|P = 1)p(X_2 = 1|P = 1)p(X_3 = 1|P = 1) = 0.8 * 0.5 * 0.4 = 0.16$ and $p(X_1 = 1|P = 0)p(X_2 = 1|P = 0)p(X_3 = 1|P = 0) = 0.5 * 0.32 * 0.16 = 0.0256$. Thus the predicted label is **positive class** $P = 1$.

Scenario B: $X_1 = 1, X_2 = 0, X_3 = 0$: We have $p(X_1 = 1|P = 1)p(X_2 = 0|P = 1)p(X_3 = 0|P = 1) = 0.8 * 0.5 * 0.6 = 0.24$ and $p(X_1 = 1|P = 0)p(X_2 = 0|P = 0)p(X_3 = 0|P = 0) = 0.5 * 0.68 * 0.84 = 0.2856$. Thus the predicted label is **negative class** $P = 0$.

Scenario C: $X_1 = 0, X_2 = 1, X_3 = 0$: We have $p(X_1 = 0|P = 1)p(X_2 = 1|P = 1)p(X_3 = 0|P = 1) = 0.2 * 0.5 * 0.6 = 0.06$ and $p(X_1 = 0|P = 0)p(X_2 = 1|P =$

$0)p(X_3 = 0|P = 0) = 0.5 * 0.32 * 0.84 = 0.1344$. Thus the predicted label is **negative class** $P = 0$.

Scenario D: $X_1 = 0, X_2 = 0, X_3 = 0$: We have $p(X_1 = 0|P = 1)p(X_2 = 0|P = 1)p(X_3 = 0|P = 1) = 0.2 * 0.5 * 0.6 = 0.06$ and $p(X_1 = 0|P = 0)p(X_2 = 0|P = 0)p(X_3 = 0|P = 0) = 0.5 * 0.68 * 0.84 = 0.2856$. Thus the predicted label is **negative class** $P = 0$.

Thus, the Nïve Bayes Classifier classifies incorrectly 8 negative instances under Scenario A, and $20 + 5 + 5$ positive instances for Scenario B-D, resulting in a training error of $38/100 = 0.38$ or 38%.

iii) Consider the directed acyclic graph shown in the figure. Determine whether each of the following independence or conditional independence assumptions are valid according to the constraints given by the graph
a) $C \perp B$.

Solution : We have

$$\begin{aligned} P(B, C) &= \sum_{\forall a} P(A = a, B, C) \\ &= \sum_{\forall a} P(B|C, A = a)P(C, A = a) = P(B) \sum_{\forall a} P(C, A = a) = P(B)P(C) \end{aligned}$$

where we used the fact that $P(B|A, C) = P(B)$. Therefore we proved that $C \perp B$.

b) $D \perp B|A$.

Solution : We have

$$\begin{aligned} P(B, D|A) &= \sum_{\forall c} P(B, D, C = c|A) = \sum_{\forall c} P(B|D, A, C = c)P(D, C = c|A) \\ &= P(B) \sum_{\forall c} P(D, C = c|A) = P(B)P(D|A) \end{aligned}$$

where we used the fact that $P(B|D, A, C) = P(B)$. Therefore we proved that $D \perp B|A$.

c) $C \perp E|B$.

Solution : We have

$$\begin{aligned}P(C,E|B) &= \sum_{\forall a} P(C,E,A=a|B) = \sum_{\forall a} P(C|E,A=a,B)P(E|A=a,B)P(A=a|B) \\ &= \sum_{\forall a} P(C|A=a,B)P(E|A=a,B)P(A=a|B) \neq P(C|B)P(E|B)\end{aligned}$$

Thus we proved that $C \not\perp E|B$, because these r.v's are both dependent on a common parent A. Note that the independence may still hold numerically but is not implied by the given Bayesian network.

iv Suppose C is the target class and we observe $A = 1, B = 1, D = 1$. Using the directed acyclic graph given in the previous question and the probability information given below, determine which is the more likely value for C (0 or 1).

Solution : To determine the more likely class value, we will need to compare the conditional probabilities $P(C = 0|A = 1, B = 1, D = 1)$ and $P(C = 1|A = 1, B = 1, D = 1)$. We have

$$P(C=0|A=1,B=1,D=1) = \frac{P(A=1,B=1,C=0,D=1)}{P(A=1,B=1,D=1)}$$

Since the denominator would be constant and equal for the two classes, we compute

$$\begin{aligned}P(A=1,B=1,C=0,D=1) &= P(A=1)P(B=1)P(C=0|A=1)P(D=1|C=0) \\ &= 0.5 * 0.6 * (1-0.7) * 0.4 = 0.036\end{aligned}$$

Similarly, we compute

$$\begin{aligned}P(A=1,B=1,C=1,D=1) &= P(A=1)P(B=1)P(C=1|A=1)P(D=1|C=1) \\ &= 0.5 * 0.6 * 0.7 * 0.3 = 0.063\end{aligned}$$

Therefore, the prediction for $A = 1, B = 1, D = 1$ would be class $C = 1$.

Remark: Since the quiz asks for the exact conditionals, we found $P(A = 1, B = 1, D = 1) = 0.5 * 0.6 * \{0.3 * 0.7 + 0.4 * 0.3\} = 0.099$ which gives $P(C = 0|A = 1, B = 1, D = 1) = 0.036/0.099 = 0.3636$ and $P(C = 1|A = 1, B = 1, D = 1) = 0.063/0.099 = 0.6363$

————————————————X——————————————————-

3. **Analyzing the kNN**

i) Given two points $x = [3\ 2.1\ 4.8\ 5.1\ 6.2]$ and $y = [9.1\ 0.7\ 2.2\ 5.1\ 1.8]$ in $\mathbb{R}^5$, compare the Euclidean and Manhattan distances.

Solution : The Euclidean distance is given by

$$d_E(x,y) = \sqrt{(3-9.1)^2 + (2.1-0.7)^2 + (4.8-2.2)^2 + (5.1-5.1)^2 + (6.2-1.8)^2} = 8.08$$

and the Manhattan distance is given by

$$d_M(x,y) = |3-9.1| + |2.1-0.7| + |4.8-2.2| + |5.1-5.1| + |6.2-1.8| = 14.5$$

Thus, the two data points are closer under the Euclidean distance.

ii) Consider a k-nearest neighbors binary classifier which assigns the class of a test point to be the class of the majority of the k-nearest neighbors, according to a Euclidean distance metric. Using the data set shown below to train the classifier and choosing k=5, what is the classification error on the training set? Assume that a point can be its own neighbor.

Solution : The training set contains 14 points in total. For k=5, the k-NN approach would wrongly classify 4 points as marked in the given figure. It can e verified that the green labels correspond to correctly classified, while the red ones correspond to misclassified data points, based on euclidean distance metric. Thus the classification error is 4/14 = 0.2857 or 28.57%.

iii) In the data set shown above, what is the value of k that minimizes the training error? Note that a point can be its own neighbor.

Solution : If a point can be its own neighbor, the training error is minimized for k=1. In fact, this is irrespective of the data set chosen.

iv) Consider a binary knn classifier where k=4 and the two labels are red and green. Consider classifying a new point $x = (1,1)$, where two of xs nearest neighbors are labeled as red and two are labeled as green as shown. Which of the following methods can be used to break ties on this dataset?

Solution : The four methods are

- Assign x the label of its nearest neighbor: In the given data set, the point x has two equidistant points as nearest neighbor. Thus, this CANNOT break a tie.
- Flip a coin to randomly assign a label to x (from the labels of its 4 closest points): Since this is a two-class problem, one CAN break a tie by the said method. However, this method may NOT BE RECOMMENDED as this virtually discards the training data set. In case of randomization, one could also use the prior class possibilities instead of flipping a coin.
- Use k=3 instead: Since in the given example, there are four equidistant points neighboring x, choosing $k = 3$ CANNOT break a tie as the set of three neighbors is not unique.
- Use k=5 instead: In the given example data set, using k=5 CAN help in breaking the tie.

v) Consider the given data concerning the relationship between academic performance and salary after graduation. Compute the salary prediction for k=3

Solution : In the given data set, we need to predict salary for the 2-D point $x = (3.5, 3.6)$. The Euclidean distances from the 8 given training points is computed below

$$\begin{aligned}
d(x, x_1)^2 &= (3.5 - 2.2)^2 + (3.6 - 3.4)^2 = 1.73 \\
d(x, x_2)^2 &= (3.5 - 3.9)^2 + (3.6 - 2.9)^2 = 0.65 \\
d(x, x_3)^2 &= (3.5 - 3.7)^2 + (3.6 - 3.6)^2 = 0.04 \\
d(x, x_4)^2 &= (3.5 - 4)^2 + (3.6 - 4)^2 = 0.41 \\
d(x, x_5)^2 &= (3.5 - 2.8)^2 + (3.6 - 3.5)^2 = 0.5 \\
d(x, x_6)^2 &= (3.5 - 3.5)^2 + (3.6 - 1.0)^2 = 6.76 \\
d(x, x_7)^2 &= (3.5 - 3.8)^2 + (3.6 - 4.0)^2 = 0.25 \\
d(x, x_8)^2 &= (3.5 - 3.1)^2 + (3.6 - 2.5)^2 = 1.37
\end{aligned}$$

Thus the 3 nearest neighbors are x_3, x_7 and x_4. Therefore, the predicted salary is $(91 + 142 + 163)/3 = 132$ K.

——————————————X——————————————-

4. **Rule-based classifier.**

Consider the following classification rule extracted from the medical history of patients:

$$\text{BloodPressure} > 150 \rightarrow \text{ HeartDisease=Severe}$$

Suppose the coverage of the rule is 5% and the accuracy is 60% on the training data. Coverage refers to the proportion of patients in the training set who satisfy the rule condition (i.e., left-hand side of the rule) while accuracy is the fraction of such patients (who satisfy the rule condition) who also have heart disease.

i) Explain whether coverage and accuracy (1) will stay the same, (2) will increase or stay the same, (3) will decrease or stay the same, or (4) can change in either direction (increase/decrease), if we add the conjunct CholesterolLevel $>$ 245, to the left-hand side of the rule.

Solution : When we add a conjunct, only fewer people can qualify the additional requirement. Thus, the coverage will decrease or remain unchanged. Out of the people removed due to the additional conjunct, some could be diseased while some may not. Therefore, the accuracy can sway both ways. Thus, coverage (3) will decrease or stay the same; and accuracy (4) can change in either direction (increase/decrease).

ii) Explain whether coverage and accuracy (1) will stay the same, (2) will increase or stay the same, (3) will decrease or stay the same, or (4) can change in either direction (increase/decrease), if the rule condition is BloodPressure $>$ 200 instead of 150.

Solution : Since the qualifying criteria of the conjunct is being made more restrictive, only fewer people can qualify. Thus, coverage (3) will decrease or stay the same. Out of the people removed due to the more restrictive conjunct, some could be diseased while some may not. Therefore, accuracy (4) can change in either direction (increase/decrease).

Comment: Instead of a 3-class problem (HeartDisease is Severe, Mild, and None), suppose we reduce this to a 2-class problem (HeartDisease is Yes or No), where all the training examples assigned to the Severe and Mild classes are in the Yes category while the remaining is in the No category.

iv) Explain whether coverage and accuracy (1) will stay the same, (2) will increase or stay the same, (3) will decrease or stay the same, or (4) can change in either direction (increase/decrease) when the rule becomes BloodPressure $>$ 150 $\rightarrow$ HeartDisease=Yes

Solution : In this case, the conjunct or the qualifying criteria remains unchanged. Thus, the coverage (1) will stay the same. However, those people qualifying for conjunct and having a Mild heart disease were earlier counted as incorrect predictions. Thus, the denominator remained unchanged and the numerator set of correct predictions became larger, unless there were no cases of Mild disease in the selected set. Therefore, the accuracy (2) will increase or stay the same.

v) The coverage and accuracy for the previous rule are computed using patients as training examples. However, the original EMR database contains records of patient visits to the healthcare provider. Assume that blood pressure is taken at every visit by a patient. The training example for a patient is created by merging all the visit records associated with that patient in the following way:

- The BloodPressure attribute is computed based on the maximum BloodPressure value ever recorded for the given patient. For example, if the BloodPressure values recorded for a patient are 129, 147, 140, and 133, the BloodPressure value for that patient in the training set is the highest value, 147. Different patients may have different maximum recorded value.
- The HeartDisease class is determined based on the number of visits related to heart-related incidents. If a patient makes at least 3 visits where the ChiefComplaint attribute value is HeartRelated, then the patient is classified as HeartDisease=Severe. Explain whether coverage and accuracy (1) will stay the same, (2) will increase or stay the same, (3) will decrease or stay the same, or (4) can change in either direction (increase/decrease) for the rule BloodPressure $> 150 \rightarrow$ ChiefComplaint=HeartRelated if each training example corresponds to a patient visit (instead of a patient).

Solution : Let there be N patients denoted as $p_1, \ldots, p_N$ and patient p_i makes v_i visits. Thus in total there have been $\sum_{i=1}^{N} v_i$ visits to the clinic. Let us now investigate a few cases.

Case 1: Say $v_i = 3$ for all i. Patient p_1 had a BloodPressure > 150 an all three visits. All other patients record BloodPressure > 150 on one of the three visits. Let us first compare coverage.

$$\text{Coverage(patient)} = \frac{1}{N}, \; \text{Coverage(visit)} = \frac{N+2}{3N}$$

N-1+3 = N+2

So, here we observe: Coverage(patient)$\leq$Coverage(visit).

Case 1a: To compare accuracies consider that patients p_2 to p_N had no heart related visits. Now, first assume p_1 had all three visits as heart related. Then, the accuracies are

$$\text{Accuracy(patient)} = \frac{1}{1} = 100\%,\ \text{Coverage(visit)} = \frac{3}{N+2}$$

Clearly, we have Accuracy(patient)≥Accuracy(visit).
Case 1b: Now, consider as before, patients p_2 to p_N had no heart related visits. However, patient p_1 had one of his three visits for a heart related reason. Then the accuracies would be

$$\text{Accuracy(patient)} = \frac{0}{1} =,\ \text{Coverage(visit)} = \frac{1}{N+2}$$

So, in this we get Accuracy(patient)<Accuracy(visit). From **Case 1**, we conclude that when we compare the **accuracies** for patients vs visits they **(4) can change in either direction (increase/decrease)**.
Case 2: Say $v_i = 6$ for all i. All patients p_i to p_N had a BloodPressure > 150 exactly three of their six visits. Now, let us compare coverage.

$$\text{Coverage(patient)} = \frac{N}{N} = 100\%,\ \text{Coverage(visit)} = \frac{3N}{6N} = 50\%$$

So here, Coverage(patient)>Coverage(visit). Combining the observations of Case-1 and 2, we conclude that for comparing **coverage** of patient based rules versus visit-based rules, they **(4) can change in either direction (increase/decrease)**.

——————————————X——————————————-

5. **Evaluation Measures.**
For the given confusion matrix and cost matrix, compute the following indices

Solution : From the given matrices, we have

$$\text{True Positive, } TP = 98$$
$$\text{False Positive, } FP = 37$$
$$\text{True Negative, } TN = 143$$
$$\text{False Negative, } FN = 20$$

We now compute the first set of required metrics:

$$\text{a. Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{98 + 143}{98 + 37 + 143 + 20} = 0.8087$$

$$\text{b. Precision} = \frac{TP}{TP + FP} = \frac{98}{98 + 37} = 0.7259$$

$$\text{c. Recall} = \frac{TP}{TP + FN} = \frac{98}{98 + 20} = 0.8305$$

$$\text{d. F-score} = \frac{2TP}{2TP + FP + FN} = \frac{2 * 98}{2 * 98 + 37 + 20} = 0.7747$$

Next, we are given the following cost coefficients

$$\text{True Positive Cost, } C_{TP} = -1$$
$$\text{False Positive Cost, } C_{FP} = 1$$
$$\text{True Negative Cost, } C_{TN} = 0$$
$$\text{False Negative Cost, } C_{FN} = 100$$

The remaining metrics are computed next:

$$\text{e. Cost} = 98 * (-1) + 37 * 1 + 143 * 0 + 20 * 100 = 1939$$

$$\text{f. Sensitivity} = \frac{TP}{TP + FN} = \frac{98}{98 + 20} = 0.8305$$

$$\text{g. Specificity} = 1 - FPR = \frac{TN}{FP + TN} = \frac{143}{37 + 143} = 0.7944$$

$$\text{h. False Positive Rate} = \frac{FP}{FP + TN} = \frac{37}{37 + 143} = 0.2056$$

——————————————X——————————————-

6. **Cross-validation.**

Given below is a strategy implemented for doing cross-validation (CV) for a classification problem with a large number of features.
1. Determine a subset of useful features so that they are highly correlated with the class labels.

2. Restricting attention to this feature subset, implement a classifier (multivariate).
3. Use CV for determining the parameters of the model and also to find the error of in prediction given by the fixed model.

To illustrate the problem better, suppose you are provided with 50 samples distributed equally among the two classes, and D = 1000 features sampled from a standard normal distribution such that these features are independent of the labels. Under the given circumstances any classifier will have error rate around 50%. But when you perform the above steps as follows: (1) pick the top 100 features when they are sorted in descending order of their correlation with the labels, and then (2) in next step make use of a 1-NN classifier that just uses those 100 features. The mean CV error rate was around 2.7% (much below the original error rate of 50%) after repeating this configuration 100 times. What do you think is the reason for your observation? Do you think we have performed CV correctly? If yes explain. If no suggest an improved alternative?

Solution : The outcome of the presented example clearly indicates that the CV was not performed correctly. We know this because the randomly generated features had no information about the label (one can verify this using mutual information for independent rv's). Therefore, no matter the design of a classifier, the test or CV accuracy should not be much different from the random guess $\sim 50\%$.

The main source of error comes from using the test data in step one for feature selection. Unless we have a trained model that is ready for evaluation, the CV data (acting as test data) should not be used. Thus, the correct method of CV in this example of feature selection would involve:

I. Using a desired CV strategy, say k-fold, separate out a test data set.
II. Using the *train* data set, find the N most relevant features using the correlation measures.
III. Train a classifier using the training data and N selected features.
IV. Test the accuracy on the CV set.
V. The CV accuracy may be used to decide on the *number of features*, N, if desired. In any case, the k-fold cross-validation should be used in deciding and fixing the most relevant features, potentially based on voting.
VI. Finally, once the desired features are fixed, a fresh set of CV shall be done to report the average error rates.