CS 5525 Assignment 3

Due 22th Oct 2020

Problem 1. Probability **7 = 2 + 5**

Let's suppose that we are given the joint probability distribution, see table I, of two random variables X, Y, where X represents the r.v "it will rain in Pittsburgh" and Y the r.v. "it will rain in Blacksburg".

TABLE I: Probability mass function.

X	Y	$p(X,Y)$	remarks
0	0	α	no rain
0	1	β	rains in Blacksburg, not in Pittsburgh
1	0	γ	rains in Pittsburgh, not in Blacksburg
1	1	δ	rains in both places

1)Find the expected values of X and Y as functions of $\alpha, \beta, \gamma, \delta$ **[2 points]**.

2)Find the necessary and sufficient conditions so that X and Y are independent **[5 points]**.

Problem 2. Bayes Theorem & Naïve Bayes Classifier **26 = 3 + 7 + 12 + 4**

1) 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. [**3 points**]

2) Consider a training set with 3 features, X_1, X_2 and X_3, for a binary classification problem. The distribution of the data set is shown in the table below. [**7 points**]

a) Based on the information above, determine whether X_1 and X_2 are independent of each other.

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

c) Compute the class conditional probabilities $P(X_1 = 1 \mid +)$, $P(X_1 = 1 \mid -)$, $P(X_2 = 1 \mid +)$, $P(X_2 = 1 \mid -)$, $P(X_3 = 1 \mid +)$, and $P(X_3 = 1 \mid -)$.

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.

TABLE II: Training data set.

X_1	X_2	X_3	Number of positive examples	Number of negative examples
1	1	1	20	8
1	0	0	20	17
0	1	0	5	8
0	0	0	5	17

3) Consider the directed acyclic graph shown in the figure below. Determine whether each of the following independence or conditional independence assumptions are valid according to the constraints given by the graph. To receive full credit, make sure you show your steps clearly (to prove/disprove the assumptions). [**12 points**]

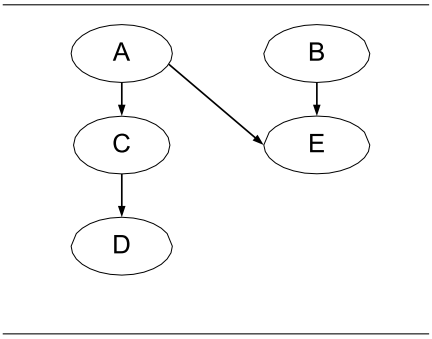


Fig. 1: Directed acyclic graph.

a) C $\perp$ B

b) D $\perp$ B | A

c) C $\perp$ E | B

4) 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). To receive full credit, you must show your steps clearly. Hint: compare $P(C = 0 \mid A = 1, B = 1, D = 1)$ against $P(C = 1 \mid A = 1, B = 1, D = 1)$. [**4 points**]

$P(A=1)=0.5$

$P(B=1)=0.6$

$P(C=1|A=0)=0.4$

$P(C=1|A=1)=0.7$

$P(D=1|C=1)=0.3$

$P(D=1|C=0)=0.4$

$P(E=1|A=1,B=1)=0.5$

$P(E=1|A=1,B=0)=0.5$

$P(E=1|A=0,B=1)=0.2$

$P(E=1|A=0,B=0)=0.4$

Problem 3. Analyzing the kNN **17 = 3 + 4 + 3 + 3 + 4**

(1) Consider the description of the two objects below:

	Object A	Object B
Feature 1	3	9.1
Feature 2	2.1	0.7
Feature 3	4.8	2.2
Feature 4	5.1	5.1
Feature 5	6.2	1.8

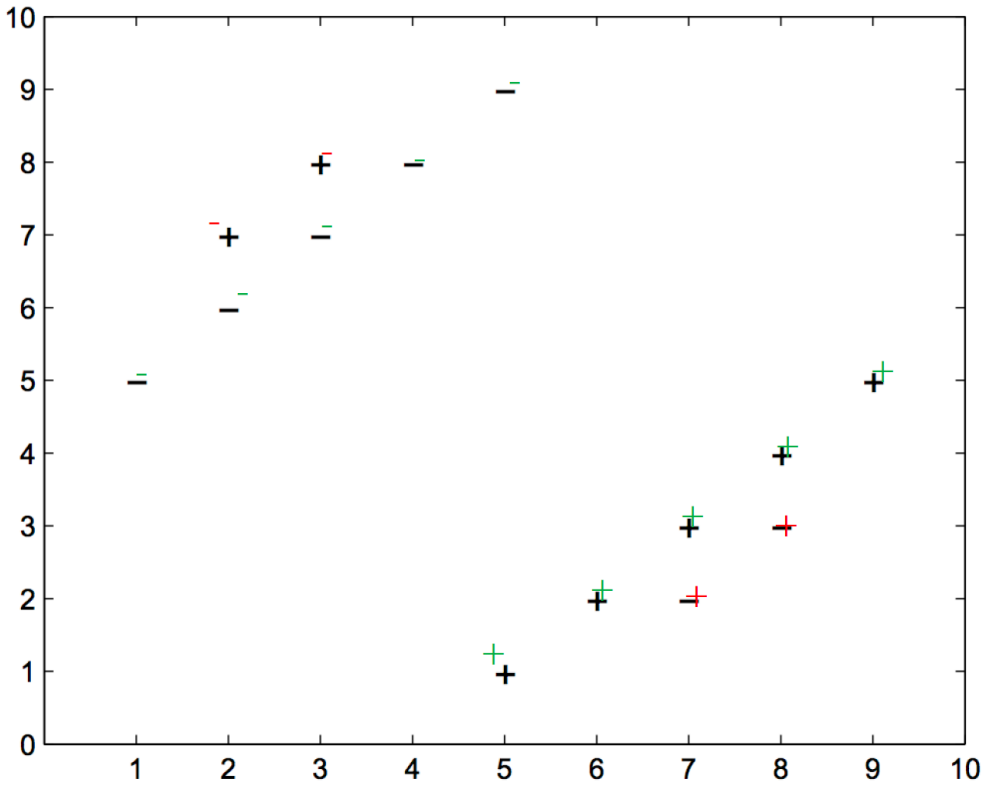
We can reason about these objects as points in high dimensional space. Consider the two different distance functions below. Under which scheme are they closer in 5-D space?

[3 points]

1) Euclidean Distance: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
2) Manhattan Distance: $d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$

(2) 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.

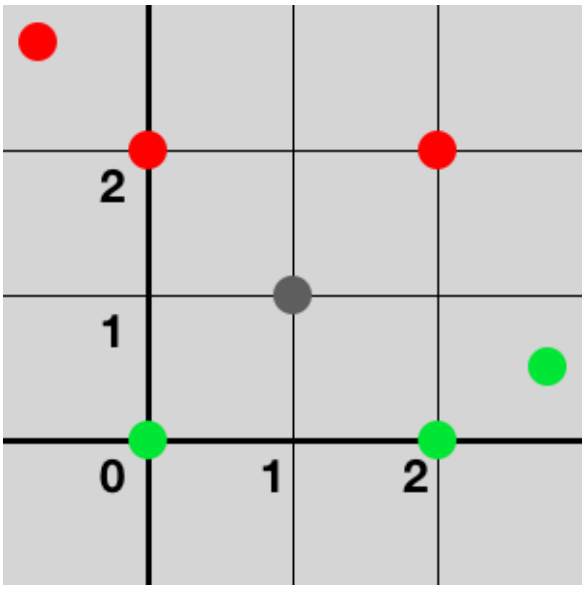
[**4 points**]



overall we have 4 errors of 14 point

(3) 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. **[3 points]**

(4) 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 x's nearest neighbors are labeled as "red" and two are labeled as "green" as shown below.



Which of the following methods can be used to break ties or avoid ties on this dataset? [**3 points**]

1) Assign x the label of its nearest neighbor

2) Flip a coin to randomly assign a label to x (from the labels of its 4 closest points)

3) Use k=3 instead

4) Use k=5 instead

(5) Consider the following data concerning the relationship between academic performance and salary after graduation. High school GPA and university GPA are two numerical variables (predictors) and salary is the numerical target. Note that salary is measured in thousands of dollars per year.

Student ID	High School GPA	University GPA	Salary
1	2.2	3.4	45
2	3.9	2.9	55
3	3.7	3.6	91
4	4.0	4.0	142
5	2.8	3.5	88
6	3.5	1.0	2600
7	3.8	4.0	163
8	3.1	2.5	67
9	3.5	3.6	unknown

In the data set shown above, our task is to predict the salary Student 9 earns after graduation. We apply kNN to this regression problem: the prediction for the numerical target (salary in this example) is equal to the average of salaries for the top k nearest neighbors.
If k=3, what is our prediction for Student 9's salary? [**4 points**]

Problem 4. Rule-based classifier. **10 = 2 + 2 + 2 + 4**

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

`BloodPressure` > 150 → `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.

1) 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. **[2 points]**
2) 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. **[2 points]**
3) 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.
4) 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 → `HeartDisease=Yes` **[2 points]**
5) 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 → `ChiefComplaint=HeartRelated` if each training example corresponds to a patient visit (instead of a patient). **[4 points]**

Problem 5. Evaluation Measures **20 = 2.5 × 8**

For the Confusion Matrix shown below, compute the following values:

Confusion Matrix	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	98	20
	-	37	143

Fig. 2: Confusion Matrix.

(a) Accuracy (b) Precision (c) Recall (d) F-measure

(e) Cost (f) Sensitivity (g) Specificity (h) False Positive Rate

For computing the cost use the matrix given in Figure 3.

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

Fig. 3: Cost Matrix.

Problem 6. Cross-validation **20 points**

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?
(Note: *Your answer should not depend on the classifier used.*)