EE514 & CS535 - Machine Learning

Mid Examination Spring 2021

Part - 2 - Solutions (40 pts)

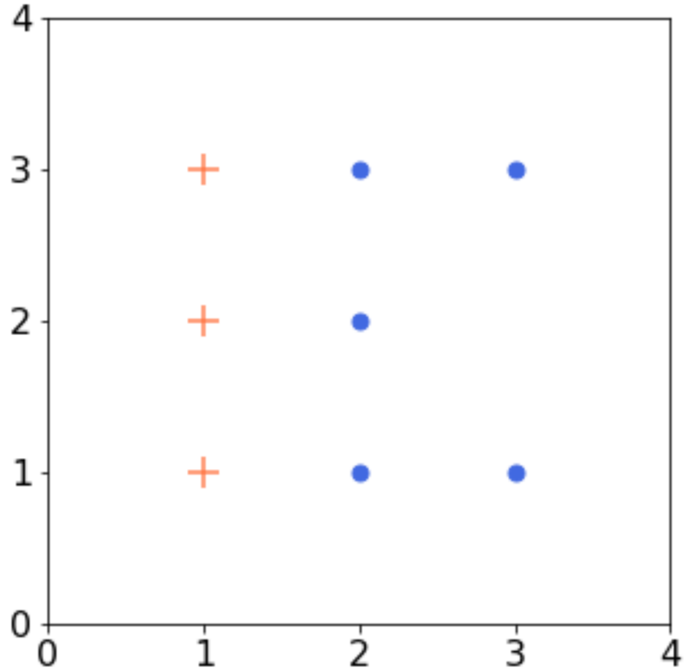
Note:

There are six questions worth 55 points. We require you to attempt questions with cumulative worth of 40 points. Do not over attempt; we will not grade the one with the highest marks.

Problem 1. (**5 pts**) For the following two dimensional labeled data, assign a label to the test point $(1.3; 4)$ using kNN for

(a) $k = 1$.

(b) $k = 3$.



Solution:

Positive (+) for $k = 1$ as the nearest neighbor is $(1, 3)$ which is positive. $\circ$ for $k = 3$ as the three nearest neighbors are $(1, 3)$, $(2, 3)$ and $(3, 3)$, the last two of which are $\circ$e.

Problem 2. (**5 pts**) In a rare collaboration, computer scientists, doctors and engineers have developed a test for detecting Examophobia disease among the students. Research suggests that

- the probability of positive outcome for the student who is suffering from this disease is 0.8.
- the probability of positive outcome for the student who is not suffering from this disease is 0.01.
- 35% of the students suffer from Examophobia.

What is the probability that the student tested positive is suffering from Examophobia?

Solution:

We employ Bayes' Theorem to determine the probability that the student tested positive is suffering from Examophobia.

$$P(+ve|\text{disease}) = 0.8 P(+ve|\text{no disease}) = 0.01 P(\text{disease}) = 0.35 P(\text{no disease}) = 1 - 0.35 = 0.6$$

Using the law of total probability, we obtain

$$P(+ve) = P(+ve|\text{disease})P(\text{disease}) + P(+ve|\text{no disease})P(\text{no disease}) = 0.2865$$

Now we use Bayes' theorem as follows

$$P(\text{disease}| + ve) = \frac{P(+ve|\text{disease})P(\text{disease})}{P(+ve)} = 0.9773$$

Problem 3. (**10 pts**)

Pakistan International Airlines has developed 2 different classifiers (A and B) for the prediction whether a flight originating from Lahore will arrive at its final destination on time or not. True or Positive here is 'On time' and it refers to the case when the flight is no more than 5 minutes late than the scheduled time. The classifiers were tested on a data-set of 500 flights, and the results are as follows:

	Actual	
	On time	Late
Classifier A, predicted on time	131	155
Classifier A, predicted late	19	195
Classifier B, predicted on time	82	72
Classifier B, predicted late	68	278

(a) Which is the preferable classifier in terms of F_1 score?

(b) Which is the preferable classifier in terms of accuracy?

Solution:

We first construct confusion matrix for both the classifiers

Classifier A

TP=131	**FP=155**
FN=19	**TN=195**

Classifier B

TP=82	**FP=72**
FN=68	**TN=278**

Noting

$$F_1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2(\text{Precision})(\text{Recall})}{(\text{Precision}) + (\text{Recall})} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

and

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}},$$

we obtain $F_1(A) = 0.6009$ and $F_1(B) = 0.5395$; $\text{Accuracy}(A) = 0.652$ and $\text{Accuracy}(B) = 0.72$, and conclude that the classifier A is better in terms of F_1 score and classifier B is better in terms of accuracy.

Problem 4. (**10 pts**) In Ridge regression, we minimize the following loss function

$$\mathcal{L}_{\text{reg}}(\mathbf{w}) = \frac{1}{2}\|(\mathbf{y} - \mathbf{X}\mathbf{w})\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

in order to find the model parameters $\mathbf{w} \in \mathbf{R}^d$. Here $\mathbf{X} \in \mathbf{R}^{n \times d}$ is the data matrix that is constructed using the features (inputs) in the training data, $\mathbf{y}$ represents the observations in the training data and $\lambda > 0$ is the regularization parameter.

(a) (**8 pts**) Derive the closed-form solution for the ridge regression problem, that is,find optimal $\mathbf{w}$ in terms of $\mathbf{X}$, $\mathbf{y}$ and λ that minimizes the loss function.

(b) (**2 pts**) How does the solution change as $\lambda \to 0$ and $\lambda \to \infty$?

Solution(a):

Noting that the loss function is convex, we take its gradient as

$$\nabla\mathcal{L}_{\text{reg}}(\mathbf{w}) = \frac{1}{2}\big(- 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w}\big)$$

and substitute it equal to zero to obtain

$$\mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = \mathbf{X}^T\mathbf{y}$$
$$\Rightarrow \big(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\big)\mathbf{w} = \mathbf{X}^T\mathbf{y}$$
$$\Rightarrow \mathbf{w} = \big(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y}$$

Solution(b):

- $\lambda = 0$, we have non-regularized solution.
- $\lambda = \infty$, the solution is a zero vector.

Problem 5. (10 pts)

(a) (**3 pts**) Draw a Bayesian network representing Naïve Bayes classifier. Assume that the d features are denoted by $x^{(1)}, x^{(2)}, \ldots, x^{(d)}$ and a class label is denoted by y.

(b) (**7 pts**) Consider a following Bayesian network. We assume that variables are boolean. Write the joint probability distribution $P(A, B, C, D, E, F, G, H, I)$ as a product of conditional distributions factored according to the Bayesian network.

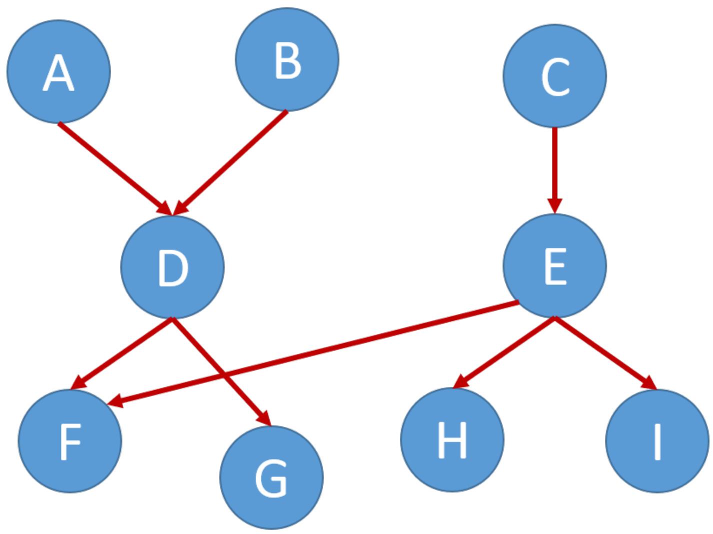
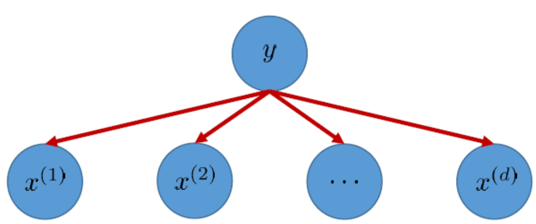


Figure 1: Bayesian Network

Solution(a):



$$P(y \mid \mathbf{x}) = \prod_{i=1}^{d} P(x^{(i)} \mid y)\, P(y)$$

Solution(b):

$$P(A, B, C, D, E, F, G, H, I) = P(A)\, P(B)\, P(C)\, P(D|A, B)\, P(E|C)\, P(F|D, E)\, P(G|D)\, P(H|E)\, P(I|E)$$

Problem 6. (**15 pts**) This problem is related to Naïve Bayes classier for text classification that help us in classifying the given statement as 'Political' or 'Not political'.

Consider the following training and test data. We will build a Naïve Bayes classifier using bag of words approach to assign labels to the last two test statements.

Category	**Statement**
Political	the election is over
Political	very clean debate
Political	the election is a match
Not Political	cricket is a great game to play
Not Political	a close but forgettable match
?	**a very close match**
?	**a very close election**

(a) (**2 pts**) Create a vocabulary for the given training data.

(b) (**4 pts**) Develop a bag of words representation for each class using the training data.

(c) (**3 pts**) Do we need to use Laplace Smoothing for the given training and test data? Provide brief justification to support your answer.

(d) (**6 pts**) Classify the given test statements as 'Political' or 'Not Political' using the Naïve Bayes approach, that is, compute $P(\text{Political})P(\text{statement} \mid \text{Political})$ and $P(\text{Not Political})P(\text{statement} \mid \text{Not Political})$ for each test statement. Use add-1 smoothing for the computation of probabilities.

Solution(a):

Vocabulary = {the, election, is, over, very, clean, debate,a, match, cricket, great, game, to, play, close, but, forgettable, match}

Solution(b):

Bag of words	Total	Political	Not Political
the	2	2	0
election	2	2	0
is	3	2	1
over	1	1	0
very	1	1	0
clean	1	1	0
debate	1	1	0
a	3	1	2
match	2	1	1
cricket	1	0	1
great	1	0	1
game	1	0	1
to	1	0	1
play	1	0	1
close	1	0	1
but	1	0	1
forgettable	1	0	1

Solution(c):

Laplace smoothing will help because we have words in the bag that are not appearing in both the categories. Otherwise, the product of probabilities will become zero and we won't be able to carry out classification.

Solution(d):

Given statement: **a very close match**

We use P and NP to refer to political and non-political respectively. We compute probabilities for each word given P.

$$P(\text{a}|P) = \frac{1+1}{12+17} = \frac{2}{29}, \quad P(\text{very}|P) = \frac{1+1}{12+17} = \frac{2}{29}$$
$$P(\text{close}|P) = \frac{0+1}{12+17} = \frac{1}{29}, \quad P(\text{match}|P) = \frac{1+1}{12+17} = \frac{2}{29}$$

Now we compute probabilities for each word given NP, that is,

$$P(\text{a}|NP) = \frac{2+1}{12+17} = \frac{3}{29}, \quad P(\text{very}|MP) = \frac{0+1}{12+17} = \frac{1}{29}$$
$$P(\text{close}|NP) = \frac{1+1}{12+17} = \frac{2}{29}, \quad P(\text{match}|NP) = \frac{1+1}{12+17} = \frac{2}{29}$$

We also note that $P(P) = \frac{3}{5}$ and $P(NP) = \frac{2}{5}$. Using these probabilities we compute

$$P(\text{a very close match}|P)P(P) = \frac{2}{29}\frac{2}{29}\frac{1}{29}\frac{2}{29}\frac{3}{5} = 6.786 \times 10^{-6}$$

and

$$P(\text{a very close match}|NP)P(NP) = \frac{3}{29}\frac{1}{29}\frac{2}{29}\frac{2}{29}\frac{2}{5} = 6.786 \times^{1} 0{-}6$$

We have same probabilities. We can either randomly classify or cannot classify given the training data.

Given statement: **a very close election**

Noting

$$P(\text{election}|P) = \frac{2+1}{12+17} = \frac{3}{29}, \quad P(\text{election}|NP) = \frac{0+1}{12+17} = \frac{1}{29}$$

Using these probabilities we compute

$$P(\text{a very close election}|P)P(P) = \frac{2}{29}\frac{2}{29}\frac{1}{29}\frac{3}{29}\frac{3}{5} = 1.01798 \times 10^{-5}$$

and

$$P(\text{a very close election}|NP)P(NP) = \frac{3}{29}\frac{1}{29}\frac{2}{29}\frac{1}{29}\frac{2}{5} = 3.393 \times^{1} 0{-}6$$

Since $P(\text{a very close election}|P)P(P) = 3P(\text{a very close election}|NP)P(NP)$, the statement is classified as 'Political'.