CS 165B – Machine Learning, Spring 2023

Assignment #1 by *Qiyuan Zhuang*
Due Thursday, April 20 by 11:59pm

Problem #1 [10 points]
For web search such as Google, explain in which parts could machine learning be used.

Answer #1:

1. **Read the input of users.** When users click "search", web need to read and understand what users have entered in search bar. It's not only a simple comparation between words and data on cloud, but much more complex because the point of users must be understood clearly. For example, if users want to search for a course "CS 165B ucsb", it cannot be understood like a person ranking 165 in a videogame(CS) competition in ucsb. Such function may be done by by techniques like NLP and models of lexical analyzing and semantic analyzing.
2. **Ranking of search results.** Obviously search results are based on the input, models and data in databases. Whereas, users will be pleased to see what they want at the beginning of search results, which means the ranking matters, thus the web can analyze what users always search and click into recently and then generate a personalized ranking of results. Meanwhile, the time and location of searching, the click rate of each page in results should also be considered.
3. **Autocorrect the input.** Sometimes users may do a typo or enter some weird phrases during searching. It will be much convenient for users if these mistakes can be detected and autocorrected. It may be done by training some frequently appearing mistakes associated with corrected data.

Problem #2 [10 points]

(a) What is the difference between classification and regression?

(b) L2 regularization is a common method to prevent the machine learning model from overfitting. It penalizes weights/coefficients with large L2 norm. Why large weight/coefficient is a typical sign of overfitting?

Answer #2:

(a) The main difference is the type of predicted output.
Classification is a process of approximating a map(function) from input variables to **discrete** output variables(labels), assigning the target variable to one of N states. Data is labelled individually as a result of input.
However, regression is a process of approximating a map from input variables to **continuous** output variables, assigning the target variable to a real-valued function of input. It's used to predict a real value of a data point.

(b) Weight/coefficient and training data determines the complexity of model. If a weight is large, it will let the model fit the related training data too well, which may lead to an overfitting. The reason is that a small change or float of training data multiplied by the corresponding coefficient is likely to cause a huge effect of result. Hence, to avoid such case, a penalty method to reduce large weights matters a lot.

Problem #3 [10 points]

You are asked to build a machine learning system to estimate someone's blood pressure (two numbers: systolic and diastolic; consider them to be real-valued) based on the following inputs: the patient's sex, age, weight, average grams of fat consumed per day, number of servings of red meat per week, servings of fruits and vegetables per day, smoker or non-smoker. You are given a training data set of values for all of these variables and the blood pressure numbers for 10,000 patients.

Answer (and explain) the following questions:

(a) What kind of machine learning problem is this?

(b) Is it a predictive task or a descriptive task?

(c) Are you likely to use a geometric model, a probabilistic model, or a logical model?

(d) Will your model be a grouping model or a grading model?

(e) What is the label space for this problem?

Answer #4:

(a) regression

(b) predictive task

(c) geometric model

(d) grading model

(e) There's no label space, since we are predicting continuous variables instead of specific labels.

Problem #4 [10 points]

Suppose you pick up a coin, it has probability p to have heads on both sides, and probability $1 - p$ to be a fair coin (i.e. have heads on one side and tails on the other side).

(a) You flip your coin once and it comes up head. What is the chance that you picked up the fair coin, given the observation?

(b) You flip your coin $\boldsymbol{n}$ times and it comes up heads for all the time. What is the chance that you picked up the fair coin, given the observation?

Answer #4:

(a) We can use **Bayes Rules**. Define X as whether the coin is fair or not. Define Y as the observation.

$$P(X = fair \mid Y = head) = \frac{P(X = fair)P(Y = head \mid X = fair)}{\sum P(X)P(Y = head \mid X)}$$
$$= \frac{0.5(1-p)}{0.5(1-p)+p} = \frac{1-p}{1+p}$$

(b) $P(X = fair \mid Y = head, \dots, Y = head)$
$$= \frac{P(X = fair)P(Y = head, \dots, Y = head \mid X = fair)}{\sum P(X)P(Y = head, \dots, Y = head \mid X)}$$
$$= \frac{0.5^n(1-p)}{0.5^n(1-p)+p} = \frac{1-p}{1+(2^n-1)p}$$

Problem #5 [18 points]

We (simplistically) characterize a planet in the solar system in terms of the following statistics: (omitted)

Treating the statistics for each player ($x1$ and $x2$) as a feature vector, what is the distance between them, measured in terms of

(a) L1 distance,

(b) L2 distance,

(c) L_{10} distance?

(d) If a constant vector $v = [50\ 2\ 0.4\ 0.5\ 0.1\ 10]^T$ is added to both $x1$ and $x2$, which (if any) of L1, L2, or L_{10} will change?

(e) If $x1$ and $x2$ are multiplied by a constant k, which (if any) of L1, L2, or L_{10} will change?

(f) What is a potential problem of using $L_{0.5}$ distance? (Hint: plot or draw $L_{0.5}$ norm on a 2D plane, how is the shape different from the other distances?)

Answer #5:

(a) $d_1 = \sum_{i=1}^{6} |x1_i - x2_i| = 130.578$

(b) $d_2 = (\sum_{i=1}^{6} (x1_i - x2_i)^2)^{\frac{1}{2}} = 89.41$

(c) $d_{10} = (\sum_{i=1}^{6} (x1_i - x2_i)^{10})^{\frac{1}{10}} = 77.0357$

(d) **None of them will change.** The reason is that difference between each element of feature vector of the two players will not change after the same vector is added to both.

(e) **All of them will change.** For each distance,

$$d' = (\sum_{i=1}^{6} k^m (x1_i - x2_i)^m)^{\frac{1}{m}} = k(\sum_{i=1}^{6} k^m (x1_i - x2_i)^m)^{\frac{1}{m}} = kd$$

(f) From the plot, we notice that $L_{0.5}$ norm grows much more rapidly than other norms. That means $L_{0.5}$ distance will be much more influenced by larger elements in feature vectors. It may lead to a completely biased result and thus unstability of model.

Problem #6 [12 points]

The joint probability distribution of three variables, class, grade and effort can be computed from the following table that shows numbers of students in each bin: (omitted)

(a) What is the conditional probability distribution P(*grade* | *class=165B, effort=Small*)?

(b) What is the marginal probability distribution P(*grade, effort*)?

(c) What is the marginal probability distribution P(*effort*)?

(d) What is P(*grade*=A | *class=basketweaving*)?

Answer #6:

(a) Grade=A: 1/30
Grade=B: 1/6
Grade=C: 3/10
Grade=D: 1/3
Grade=F: 1/6

(b)

	A	B	C	D	F
Small	0.055	0.1	0.07	0.075	0.025
Medium	0.145	0.1	0.05	0.02	0
Large	0.25	0.1	0.005	0.005	0

(c) Effort=small: 0.325
Effort=medium: 0.315
Effort=large: 0.36

(d) P=300/500=0.6

Problem #7 [20 points]

There are 10,000 images used to train a fine-grained dog classification system – 1,000 of them are Shiba, and the rest are Husky. To test the system, you have 1,000 images – 400 Shiba and 600 Husky – in your test set.

The results of the test are as follows: 150 Shibas are mis-classified as Husky and the rest are classified correctly; 50 Huskies are mis-classified as Shiba and the rest are classified correctly.

(a) Show the contingency table for this binary classification experiment. Label it clearly and fill out the table entries.

(b) What is the false positive rate of the system in this experiment?

(c) What is the false negative rate?

(d) What is the error rate and accuracy?

(e) What is the precision?

(f) What is the problem with the training set? Propose a solution and clearly state any assumption you make.

Answer #7:

(a) Define Husky as positive one.

	Husky	Shiba
Estimated Husky	550(TP)	150(FP)
Estimated Shiba	50(FN)	250(TN)

(b) $FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = \frac{3}{8}$

(c) $FNR = \frac{FN}{P} = \frac{FN}{FN+TP} = \frac{1}{12}$

(d) Error rate$= \frac{FP+FN}{P+N} = 0.2$

Accuracy$= \frac{TP+TN}{P+N} = 0.8$

(e) Precision$= \frac{TP}{\hat{P}} = \frac{11}{14}$

(f) There are too many Huskies in the training set, while the number of Shibas is too small.

Solu. We can increase the number of Shibas in the training set.(at least 40% of total)

Assump. Every data point in training set is correctly labelled.

Problem #8 [10 points]

Are these four samples in the below picture linearly separable? Consider the feature transformation $\varphi(x) = [1, x_1^2, x_2^2, \sqrt{2}x_1x_2]$, are these four samples linearly separable in the feature space $\varphi(x)$? Explain why it is or not linearly separable.

Answer #8:

(a) Samples in picture is **not** linearly separable.
Notice the equation of the line joining two "+" points is y-x=0.
For the "-" on the top, y-x=2>0, while for the "-" below, y-x=-2<0.
Hence, They are **not** linearly separable.

(b) Firstly, we obtain the new four points:

+1: (1, 1, 1, $\sqrt{2}$)

+2: (1, 1, 1, $\sqrt{2}$)

-1: (1, 1, 1, -$\sqrt{2}$)

-2: (1, 1, 1, -$\sqrt{2}$)

Obviously they are linearly separable, since the position of +1 and +2 is the same(so as -1 and -2). So we can regard these 4 points as two points.
And there must be a 4D line to separate two different points. Proved.