CS 165B – Machine Learning, Spring 2023

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

Notes:

- *This assignment is to be done individually. You may discuss the problems at a general level with others in the class (e.g., about the concepts underlying the question, or what lecture or reading material may be relevant), but the work you turn in must be solely your own.*
- Be sure to re-read the "Policy on Academic Integrity" on the course syllabus.
- Be aware of the late policy in the course syllabus – i.e., *20% grade reduction every day for homework and machine problems*.
- Justify every answer you give – **show the work** that achieves the answer or **explain** your response.
- Any updates or corrections will be posted on Piazza, so check there occasionally.
- **All assignments must be clear and legible.** It is recommended to type the solutions on this Microsoft Word file directly and save it as PDF. If you'll be submitting a handwritten assignment, please ensure that it's readable and neat. If your writing is not easily readable, your solution is not easy to follow on the page, or your PDF is not of very high quality, your assignment will not be graded. DO NOT submit picture of your written work. (If you must scan your written work in, use a high-quality scanner. Plan in advance.)
- To turn in your assignment:
  - Register as a student for **CS 165B Spring 2023 Machine Learning** on Gradescope.
  - Submit a typeset PDF version to Gradescope

Problem #1 [5 points]
For web search such as Google, explain in which parts could machine learning be used.

Answer:
The answer is subjective. Possible applications include data mining, recommendation system, voice recognition and so on.

Problem #2 [6 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:

(a) The output of classification task is a discrete set of labels, while that of regression is continuous.

(b) Large weights mean that even a small change/perturbation in the input values can lead to a drastic change in the decision hyperplane, making the model fit to every single sample of the training set. (refer to the overfitting plot in the lecture slides)

Problem #3 [6 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:

(a) They could call this supervised learning, regression (actually, two regression problems), offline learning. It is not a classification problem, however.

(b) Predictive, since we are asked to estimate the blood pressure given the features.

(c) A geometric model. For example, linear regression can be used. The others don't make sense for this problem.

(d) Grading, since the model need to predict the real value of blood pressure and the instance space is considered as a whole.

(e) Real values (with dimension of 2)

Problem #4 [6 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:

(a) $P(fair\ coin|\ t = heads) = \frac{P(fair\ coin, t=heads)}{P(heads)} = \frac{(1-p)\times 0.5}{p\times 1+(1-p)\times 0.5} = \frac{1-p}{1+p}$

(b) $P(fair\ coin|\ t_1 = t_2, \dots, t_n = heads) = \frac{P(fair\ coin, all\ heads)}{P(heads)} = \frac{(1-p)\times 0.5^n}{p\times 1+(1-p)\times 0.5^n} = \frac{1-p}{1+(2^n-1)\times p}$

Problem #5 [8 points]

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

	Earth ($x1$)	Mars ($x2$)
Distance from Sun (Billion km)	0.149	0.227
Surface Gravity (m/s^2)	9.80	3.71
Average Temperature (°C)	14	-63
Density (g/cm^3)	5.5	3.94
Volume (Earth=1)	1	0.150
Black-body Temperature (K)	254	209

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:

(a) 130.578

(b) 89.41

(c) 77.04

(d) None of them will change since the absolute different of x1 and x2 will remain the same.

(e) If k=1, they won't change, while they will change in other cases (k≠1).

(f) When the dimensionality of feature vector is small, the relative distance between samples calculated by $L_{0.5}$ is very similar, which leads to degraded results of downstream tasks like clustering. ($L_{0.5}$ is non-convex. Therefore, it is not guaranteed to converge to the global optimal with descent methods.)

Problem #6 [10 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:

	class = 165B			**class** = basketweaving		
grade	**effort**=Small	Medium	Large	**effort**=Small	Medium	Large
A	5	45	100	50	100	150
B	25	75	75	75	25	25
C	45	25	5	25	25	0
D	50	20	5	25	0	0
F	25	0	0	0	0	0

(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:

(a) P(A|165B, Small) = 5/(5+25+45+50+25) = 1/30
P(B|165B, Small) = 25/150 = 1/6
P(C|165B, Small) = 45/150 = 3/10
P(D|165B, Small) = 50/150 = 1/3
P(F|165B, Small) = 25/150 = 1/6

(b)

	Small	Medium	Large
A	55/1000=11/200=0.055	145/1000=29/200=0.145	250/1000=1/4=0.25
B	100/1000=1/10=0.1	100/1000=0.1	100/1000=0.1
C	70/1000=0.07	50/1000=0.05	5/1000=0.005

D	75/1000=0.075	20/1000=0.02	5/1000=0.005
F	25/1000=0.025	0	0

(c) P(Small) = 325/1000=0.325
P(Medium) = 315/1000 = 0.315
P(Large) = 360/1000 = 0.36
(d) P(A|Basketweaving) = 300/500 = 0.6

Problem #7 [10 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.

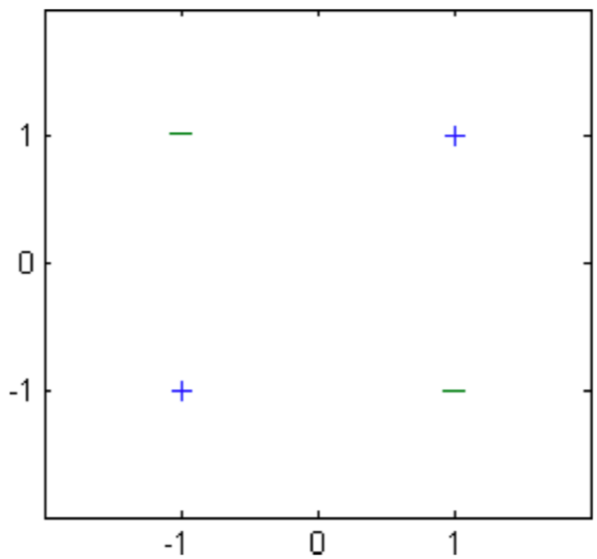
Answer:
(a)

Actual \ Predicted	Shiba	Husky	Total
Shiba	250	50	300
Husky	150	550	700
Total	400	600	1,000

(b) Use Shiba as positive: 50 / (50+550) = 1/12
Use Husky as positive: 150 / (150+250) = 3/8
(c) Use Shiba as positive: 150 / (150+250) = 3/8
Use Husky as positive: 50 / (50+550) = 1/12
(d) Error rate = 200/1000 = 0.2, accuracy = 0.8
(e) Use Shiba as positive: 250 / 300 = 5/6
Use Husky as positive: 550 / 700 = 11/14

(f) The training set is imbalanced and biased towards Husky. Possible solutions: Downsampling husky images or collect more shiba images; Use all the images but assign higher weights to Shiba images, and etc.

Problem #8 [10 points]
Are these four samples in the below picture linearly separable? Consider the feature transformation $\varphi(x) = \left[1, x_1^2, x_2^2, \sqrt{2}x_1x_2\right]$, are these four samples linearly separable in the feature space $\varphi(x)$? Explain why it is or not linearly separable.



As always, don't just give the answer – explain how you arrived at the answer.

Answer:

(a) Not linearly separable because it is impossible to find a straight line to separate them.

(b) The transformed features are:

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

We can find a coefficient $w = (0,0,0,1/\sqrt{2})$ such that $w^T x$ can be separated by a threshold 0:

$w^T x =$

$-: -1, -1$

$+: +1, +1$

Therefore, it is linearly separable with $w = (0,0,0,1/\sqrt{2})$