CS4780 Midterm

Fall 2018

NAME:	
Net ID:	
Email:	

I promise to abide by Cornell's Code of Academic Integrity.

Signature: ____________________

1 [??] General Machine Learning

Please identify if these statements are either True or False. Please justify your answer **if false**. Correct "True" questions yield 1 point. Correct "False" questions yield 2 points, one for the answer and one for the justification.

1. **(T/F)** As $n \to \infty$, the 1-NN error is no more than twice the error of the Bayes Optimal classifier.
   T

2. **(T/F)** MLE can overfit the data if n (the number of trainig samples) is small. It tends to work well when n is large.
   T.

3. **(T/F)** Both, Gradient descent and Newton's method use only a 1st order approximation of the function to be minimized.
   F. Newton's method uses 2nd order approximation of the function

4. **(T/F)** If a data set is linearly separable, the Perceptron guarantees that you find a hyperplane but the SVM finds the maximum margin separating hyperplane.
   T

5. **(T/F)** The best machine learning algorithm make no assumptions about the data.
   F. ML algorithms always make assumptions about the data.

6. **(T/F)** The k-NN classifier is not a linear classifier. T

7. **(T/F)** The k-NN algorithm can be used for classification, but not regression.
   F, k-NN can be used for regression by averaging the labels of the k nearest neighbors.

8. **(T/F)** The order of the training points can affect the training time of the Perceptron algorithm. T

9. **(T/F)** Even on non-linearly-separable datasets, the Perceptron algorithm is guaranteed to converge in finite time.
   F. For datasets that are not linearly separable, the Perceptron algorithm can never finishL: it runs forever.

10. **(T/F)** In MAP, we find the maximizer of the posterior, so we need to find an expression for the posterior.
    F. Because, $\arg\max_\theta P(\theta|D) = \arg\max_\theta \frac{P(D|\theta)P(\theta)}{P(D)} = \arg\max_\theta P(D|\theta)P(\theta)$

11. **(T/F)** If you were to use the "true" Bayesian way of machine learning you would put a prior over the possible models and draw several modelsr randomly during training.
    T

12. **(T/F)** If the features are probabilistically dependent on each other, then the naive Bayes assumption cannot hold. F. the features could be **conditionally** independent, given the label.

13. (**T/F**) Logistic regression is a generative model. F. It's a discriminative model, not a generative model.

14. (**T/F**) The order of the training points can affect the convergence of the gradient descent algorithm.
F, gradient descent just depends on a sum across the training examples, and sums are independent of order.

15. (**T/F**) For gradient descent, higher learning rates guarantee faster convergence times.
F, higher learning rates can lead to divergence.

16. (**T/F**) For Adagrad, we use the same learning rate for all features.
F, Adagrad uses different automatically-chosen learning rates for each feature.

2 [16] K-NN

In the lecture, we learn that K-NN algorithm is a distance-based algorithm. Consider that if we have different distance metric, will we get the different output of K-NN algorithm given the same data.
Suppose we have following 2D dataset:

- Class $+1$ (blue): $\{(1,5)\}$
- Class -1 (yellow): $\{(4,4),(4,0)\}$

In this problem, we will study the difference between l_2 distance and Manhattan distance. For two points $\mathbf{x} = (x_1, x_2)$ and $\mathbf{z} = (z_1, z_2)$, l_2 distance is defined by

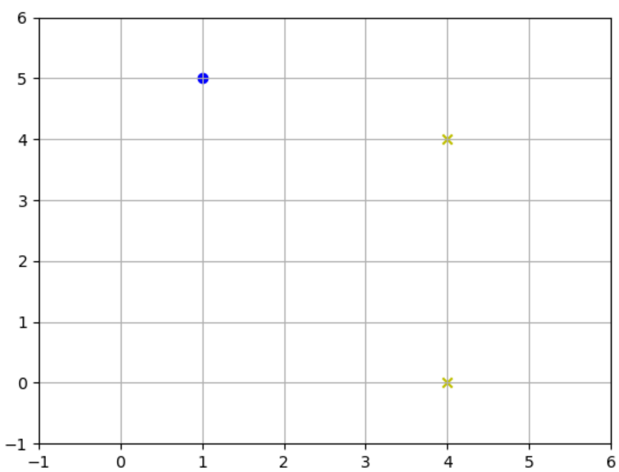
$$\mathrm{d}_1(\mathbf{x}, \mathbf{z}) = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2}, \tag{1}$$

and Manhattan distance is defined by

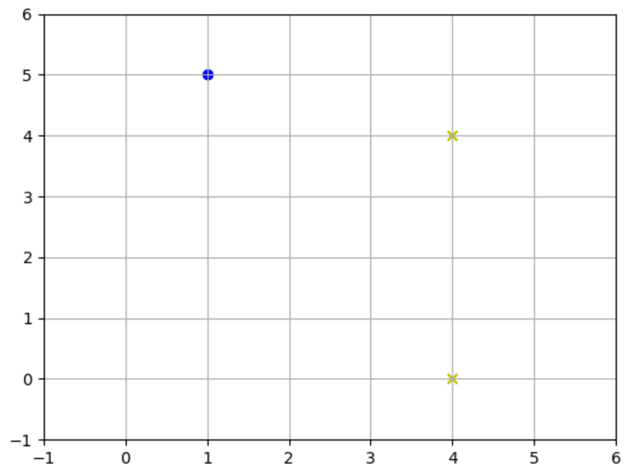
$$\mathrm{d}_2(\mathbf{x}, \mathbf{z}) = |x_1 - z_1| + |x_2 - z_2|. \tag{2}$$

1. (4 pts) How will points $(1, \frac{3}{2})$ be classified when we use l_2 distance and the 1-NN classifier? If we use manhattan distance instead, will $(1, \frac{3}{2})$ be classified in the other class? Compute the distance between from $(1, \frac{3}{2})$ to those dataset with two different distance metrics and answer the questions.

2. (6 pts) Draw the decision boundary for the 1-NN classifier with l_2 distance.



3. (6 pts) Draw the decision boundary for the 1-NN classifier with Manhattan distance.



3 [16] Perception and SVM

4 [16] Maximum Likelihood Estimation

1. (6 pts) One observation x_0 is taken on a discrete random variable with probability mass function $f(x|\theta)$, where $\theta \in \{1, 2, 3\}$. Find the MLE of θ according to different x_0 and fill the blank in Table 2.

x	$f(x\|1)$	$f(x\|2)$	$f(x\|3)$
0	$\frac{1}{3}$	$\frac{1}{4}$	0
1	$\frac{1}{3}$	$\frac{1}{4}$	0
2	0	$\frac{1}{4}$	$\frac{1}{4}$
3	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{2}$
4	$\frac{1}{6}$	0	$\frac{1}{4}$

Table 1: Probability Mass Function $f(x|\theta)$

x_0	0	1	2	3	4
MLE of θ					

Table 2: MLE Respect to x_0

2. (10 pts) Let $x_1, \cdots, x_n$ be iid random samples from the pdf

$$f(x|\theta) = \theta x^{-2} \qquad ,0 < \theta \leq x < \infty \tag{3}$$

   (a) (4 pts) Write the the likelihood function $L(\theta|x_1, \cdots, x_n)$. (Hint: it is a function of $\min_i x_i$)

   (b) (4 pts) Compute the MLE $\hat{\theta}$.

   (c) (2 pts) Consider a specific case that $n = 5$ and these five x_i are $3, 10, 6, 8, 4$ respectively. What is the MLE $\hat{\theta}$ in this case.

5 [16] Naive Bayes

1.

6 [16] Gradient Descent

In this problem, we will see Gradient Descent can minimize the loss function

$$l(w) = (w - x)^2. \tag{4}$$

1. (4 pts) Suppose at time t, we have x_t. Write the update formula for x_{t+1} using Gradient Descent when learning rate $r < 1$.

2. (6 pts) Notice that $\arg\min_x l(x) = a$. Prove that $\lim_{t \to \infty} |x_t - a| = 0$ for arbitrary starting points x_0.

3. (6 pts)Find an example of loss function $l(x)$, learning rate $r < 1$ such that $\exists x_0 \; l(x_1) > l(x_0)$, where x_1 is updated from x_0 by Gradient Descent.

True/False	
kNN	
Perception & SVM	
MLE	
NB	
Linear Classifier	
Gradient Descent	
TOTAL	