UBC CPSC 340 2017W1
MIDTERM EXAM

TIME: 55 minutes

Name: ______________________________

Student number: ______________________________

CWL username: ______________________________

Signature: ______________________________

By signing above, I hereby acknowledge that I did not / will not cheat on this exam.

- Do not open the exam until you are directed to do so.
- Once you open the exam, make sure that it contains this cover page plus 9 pages of exam questions.
- One letter-size sheet (both sides) of notes is allowed. No other material or accessories may be used.
- You may use either pen or pencil, but exams written in pencil may not be eligible for regrading.
- Please be prepared to present, upon request, a student card for identification.
- If you need more space, use the blank page at the end of the exam, and clearly indicate that your work continues there.
- Most questions require a short answer. Work efficiently and avoid writing lengthy answers.
- Unless otherwise stated, n refers to the number of training examples and d is the number of features.
- If anything is unclear or seems ambiguous, **state your assumptions**.
- Please look up occassionally in case there are clarifications writte on the projection.
- You are weclome to (quietly) leave early if you finish in under 50 minutes, but please *do not leave in the last 5 minutes* as it is very distracting to those who want to work up to the last minute.

Question:	1	2	3	4	5	Total
Points:	24	5	4	5	7	45
Score:						

G O O D L U C K !!

Question 1. **(24 points)**

Answer the questions below. *Be concise: avoid spending valuable time on lengthy answers.*

2 pts (a) What does x_{ij} refer to in the notation we've been using in class?

2 pts (b) Why shouldn't you use the training error to choose the value of k in k-nearest neighbours?

2 pts (c) What is the difference between a validation error and the test error?

2 pts (d) What is the effect of the number of features d that our model uses on the two parts of the fundamental trade-off?

2 pts (e) Explain why a random forest based on random trees of depth 10 could be viewed as a parametric classifier. Explain why or why not it would be a parametric classifier if we set the depth to ∞ in our code?

2 pts (f) What is a disadvantage of using scatterplots as a method for outlier detection?

2 pts (g) Besides finding a clustering of the data, what is another use of the k-means algorithm?

2 pts (h) What is wrong with using the code below for computing the validation error of a regression model on t examples?

```
sum(yhat .!= y)/t
```

2 pts (i) Describe a situation where it could be better to use gradient descent than the normal equations to solve a least squares problem.

2 pts (j) In regression, what is a situation where we would want to minimize the L1-norm error ($\|Xw - y\|_1$) instead of the least squares error ($\|Xw - y\|^2$)?

2 pts (k) Why would we want to approximate the L∞-norm error with the log-sum-exp function?

2 pts (l) Suppose that a famous person in the machine learning community is advertising their "extremely-deep convolutional fuzzy-genetic Hilbert-long-short recurrent neural network" classifier, which has 500 hyper-parameters. This person claims that if you take 10 different famous (and very-difficult) datasets, and tune the 500 hyper-parameters based on each dataset's validation set, that you can beat the current best-known validation set error on all 10 datasets. Explain whether or not this amazing claim is likely to be meaningful.

Question 2. **(5 points)**

Consider the dataset below, which has 10 training examples, 2 features, and 3 classes:

$$X = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix}.$$

3 pts

(a) What is a decision stump that minimizes the classification error? Briefly justify your choice.

2 pts

(b) How would your decision stump from part (a) classify the test example below?

$$\hat{x} = [0\ 0]$$

Question 3. **(4 points)**

Consider the dataset below, which has 5 training examples and 1 feature:

$$X = \begin{bmatrix} 5 \\ 3 \\ 4 \\ 2 \\ 1 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ -0.5 \\ -0.2 \\ 0 \\ -0.1 \end{bmatrix}.$$

2 pts (a) Suppose we want to fit a degree $p = 2$ polynomial to this dataset. Write a feature matrix Z that we could we use in a linear regression model to fit such a quadratic model?

2 pts (b) If we fit the data set using our standard polynomial basis with $p = 2$ and obtained the regression weights

$$w = \begin{bmatrix} 2 \\ -3 \\ 0.5 \end{bmatrix},$$

what value of y_i would we predict for the test example $\hat{x} = [2]$?

Question 4. **(5 points)**

3 pts (a) What is the time complexity of clustering using the k-medians algorithm? Answer using big-O notation with a brief explanation. Your answer may depend on n, d k, and/or the number of iterations T. You should assume a straightforward implementation of k-medians as in the assignment code.

2 pts (b) Using the same conventions as part (a), what is the cost of clustering t new objects using a trained k-medians model?

Question 5. **(7 points)**

4 pts (a) The *tilted* least squares objective function has the form

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2 + \sum_{j=1}^{d} w_j v_j,$$

for a vector v with real-valued elements v_j (you can use V as a diagonal matrix with the v_i values along the diagonal, if you need it).

Show how the minimizer of this (convex) function can be written as the solution of a linear system, explaining your steps.

3 pts (b) Consider the weighted absolute error with a penalty on the largest regression weight,

$$f(w) = \sum_{i=1}^{n} v_i |w^T x_i - y_i| + \lambda \max_j |w_j|,$$

where the max is taken over j in $\{1, 2, \ldots, d\}$, where $v_i \geq 0$ for all i, and where $\lambda \geq 0$. Write this objective in matrix and norm notation.

This page is intentionally blank. You can use it for scratch work or to continue an answer if you run out of space somewhere.