CPSC 340, 2017W1: SOLUTIONS TO MIDTERM EXAM

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?

> **Solution:** Value of feature j in training example i.

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

> **Solution:** The training error decreases as k decreases, so this will always choose $k = 1$.

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

> **Solution:** Validation error is the error you get on part of your training data that is set aside during training. The test error is the error you get on a new dataset that doesn't influence the training in any way (and where you might not have the labels).

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?

> **Solution:** As d increases the training error goes down but the approximation error goes up.

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?

> **Solution:** We can store a depth-10 decision tree in a fixed amount of memory, no matter how large n gets. If you don't restrict the depth, it's non-parametric since the trees continue to get deeper as you get more data.

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

> **Solution:** You can only look at two variables at a time.

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

> **Solution:** Finding the means as representatives for the group (as in vector quantization). Or you could want to find the cluster of a new example, or the training examples that are related to a new examples.

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
```

Solution: You can have a great regression model that has an error of 1 under this measure, and the regression model that does the best under this measure might be terrible. The reason is that noise (or numerical issues) means you usually can't exactly fit the y values.

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.

Solution: If d is large. (And gradient descent doesn't take too many iterations.)

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 ($\|Xw - y\|^2$)?

Solution: If you have outliers. (L1-norm is less sensitive to them.)

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

Solution: It's smooth, so we can minimize it with gradient descent.

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.

Solution: It's probably not meaningful. They didn't attend Lecture 4 so they don't know about optimization bias (and are probably just overfitting the validation sets).

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.

Solution: The baseline rule is assign everything to class 3, which gets an error of 6/10.
If we split on the first variable then:

- If $x_1 = 0$ we should predict 1, which gives an error 1/3 times.
- If $x_1 = 1$ we should predict 3, which gives an error 3/7 times.

This gives an error rate of 4/10.
If we split on the second variable then:

- If $x_2 = 0$ then we should predict 3, which gives an error 3/6 times.
- If $x_2 = 1$ we should predict 2, which gives an error 2/4 times.

This gives an error rate of 5/10.
So we should split on x_1 and use a rule like: if $x_1 = 0$ the predict 1 and otherwise predict 3.
(There are a number of equivalent rules.)

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

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

Solution: Since $x_1 = 0$ we should predict 1 (while in the training data this was assigned to 2).

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?

Solution:

$$Z = \begin{bmatrix} 1 & 5 & 25 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 2 & 4 \\ 1 & 1 & 1 \end{bmatrix}.$$

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 $\hat{x} = [2]$?

Solution:

$$2 \cdot 1 - 3 \cdot 2 + 0.5 \cdot 2^2 = 2 - 6 + 2 = -2.$$

It's also ok if they reverse the order and use:

$$2 \cdot 2^2 - 3 \cdot 2 + 0.5 \cdot 1 = 8 - 6 + .5 = 2.5.$$

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.

Solution: In each iteration you need to:

- Compute the L1-norm distance of each example to each mean. This is $O(nk)$ distance calculations, each costing $O(d)$, giving $O(ndk)$.
- Update the median of each cluster for each dimension. Doing this for one cluster costs $O(n)$ for each dimension (using a select algorithm) and we need to do this for d dimensions. This can be done in $O(nd)$ since each example is only assigned to one cluster, which is also in $O(ndk)$.

We need to run this for T iterations, so the total cost is $O(ndkT)$.

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?

Solution: We need to compute the L1-norm distance from each of the t examples to each of the k means, and each distance calculation costs $O(d)$. So the cost if $O(tdk)$.

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.

Solution: In matrix notation we have

$$f(w) = \frac{1}{2}\|Xw - y\|^2 + v^T w,$$

and that

$$\nabla f(w) = X^T Xw - X^T y + v.$$

Setting this equal to 0 gives the linear system

$$(X^T X)w = X^T y - v.$$

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.

Solution:

$$f(w) = \|V(Xw - y)\|_1 + \lambda\|w\|_\infty.$$