Midterm Test: Questions and Solutions

Name:

Student Number:

Closed book. 1-page, single-sided cheat sheet allowed (8.5x11in, no more than 6000 characters, 12pt font or larger if typed). No other aids allowed.

50 minutes. 9 pages. 4 questions. 76 points. Write all answers on the test booklet, using the backs of pages if necessary. The last page is blank. Clear, concise answers will receive more marks than long, rambling ones. Unless specified otherwise, all answers should be justified. Good luck!

I don’t know policy: If you do not know the answer to a question (or part), and you write “I don’t know”, you will receive 20% of the marks of that question (or part). If you just leave a question blank with no such statement, you will get 0 marks for that question.
1. (20 points total) True or False.
   For each of the following statements, state whether it is true or false without giving an explanation (2 points each):
   (a) If a function overfits data, then the training error is too small and the test error is too large.
      ANSWER: TRUE
   (b) In regularized linear regression, the regularization term increases the size of the weights, $w_1, ..., w_m$.
      ANSWER: FALSE
   (c) Cross validation produces a better estimate of validation error than simple validation.
      ANSWER: TRUE
   (d) In regularized linear regression, the optimal solution minimizes the validation error.
      ANSWER: TRUE
   (e) Logistic regression is better than nearest neighbours at fitting multi-modal data.
      ANSWER: FALSE
   (f) Logistic regression is a good choice when the data are linearly separable.
      ANSWER: TRUE
   (g) In multi-class logistic regression, we learn a linear function for each class.
      ANSWER: TRUE
   (h) Learning to recognize hand-written digits is an example of binary classification.
      ANSWER: FALSE
   (i) Linear least-squares regression tries to find values of $w$ and $w_0$ that minimize the value of $\sum_n [t^{(n)} - (w^T x^{(n)} + w_0)^2]$ where the sum is over the training data.
      ANSWER: FALSE
   (j) In nearest neighbours, learning from training data is slow, but classification of test data is fast.
      ANSWER: FALSE
2. (18 points total) *Numpy Programming*

Suppose $A$ is a $100 \times 1000$ matrix, $x$ is a 100-dimensional (rank 2) column vector, and $y$ is a 1000-dimensional (rank 1) vector. Solve each problem below using a single line of Python code and no loops (3 points each). You may assume that your program begins with the following line:

```python
import numpy as np
```

(a) Reshape $A$ into a $10 \times 100 \times 100$ array. Call the result $B$. $A$ itself does not change.

**ANSWER:** $B = \text{np.reshape}(A,[10,100,100])$

(b) Add $x$ to each column of $A$. Call the result $C$. $A$ itself does not change.

**ANSWER:** $C = A+x$

(c) Add $y$ to row 3 of $A$ and assign the result to row 37 of $A$. (Only row 37 changes.)

**ANSWER:** $A[37,:] = A[3,:] + y$

(d) Compute the sum of each row of $A$.

**ANSWER:** $\text{np.sum}(A,\text{axis}=1)$

(e) Print rows 12, 22, 32 and 42 of matrix $A$.

**ANSWER:** print $A[[12,22,32,42],:]$

(f) Compute the logarithm of each element of $A^T \cdot A \cdot y$, where $\cdot$ denotes matrix multiplication, and $^T$ denotes matrix transpose.

**ANSWER:** $\text{np.log}(\text{np.matmul}(\text{np.matmul}(A.T,A),y))$
3. (19 points total) *Linear and Logistic Regression.*

This question includes two figures, each displaying a set of 1-dimensional training data. Each point in a figure represents a training pair \((x, t)\), where \(x\) is the input value and \(t\) is the target value. \(x\) is a real number, and \(t\) is either 0 or 1.

(a) (6 points) Figure 1 shows a training set. Draw a plausible function that would be fitted to the data by linear least-squares regression. Also, draw a plausible function that would be fitted to the data by logistic regression. Clearly identify each function.

(b) (2 points) If we use these functions for binary classification, indicate the resulting decision boundaries on the figure. Clearly identify each boundary.
(c) (6 points) Figure 2 shows a training set. Draw a plausible function that would be fitted to the data by linear least-squares regression. Also, draw a plausible function that would be fitted to the data by logistic regression. Clearly identify each function.

(d) (2 points) If we use these functions for binary classification, indicate the resulting decision boundaries on the figure. Clearly identify each boundary.

(e) (3 points) In 20 words or less, what are the implications of parts (a) to (d) when using linear regression and logistic regression for classification? (Only the first 20 words of your answer will be graded.)

ANSWER: Linear regression is more sensitive to outliers than logistic regression.
4. (19 points total) We are given training data \((x^{(1)}, t^{(1)}), \ldots, (x^{(N)}, t^{(N)})\) where each input, \(x^{(n)}\), is a vector, and each target value, \(t^{(n)}\), is a real number. In addition, our learning problem has the following loss function:

\[
l(w) = \frac{1}{2} \sum_n [t^{(n)} - w^T x^{(n)}]^2
\]

where the sum is over the training data, and \(w\) is a weight vector we are trying to learn. To minimize the loss, it is convenient to let \(X\) be a matrix whose \(n^{th}\) row is \(x^{(n)}\), and let \(T\) be a column vector whose \(n^{th}\) element is \(t^{(n)}\). Prove each of the following equations:

(a) (3 points) \(w^T x^{(n)} = [Xw]_n\)

**ANSWER:** First note that \(X_{nm} = x^{(n)}_m\). Thus,

\[
w^T x^{(n)} = \sum_m w_m x^{(n)}_m \\
= \sum_m w_m X_{nm} \\
= [Xw]_n \quad \text{by the definition of matrix-vector multiplication.}
\]

(b) (4 points) \(\partial l(w)/\partial w_m = -\sum_n [t^{(n)} - w^T x^{(n)}] x^{(n)}_m\)

**ANSWER:** Note that \(\partial w^T x/\partial w_m = x_m\). Thus,

\[
\partial l(w)/\partial w_m = \frac{1}{2} \sum_n \partial [t^{(n)} - w^T x^{(n)}]^2/\partial w_m \\
= -\sum_n [t^{(n)} - w^T x^{(n)}] \partial w^T x^{(n)}/\partial w_m \quad \text{by the chain rule} \\
= -\sum_n [t^{(n)} - w^T x^{(n)}] x^{(n)}_m \quad \text{by the note above}
\]
(c) (12 points) $\partial l(w)/\partial w = X^T(Xw - T)$

ANSWER: The $m^{th}$ component of $\partial l(w)/\partial w$ is given by

$$\left[\partial l(w)/\partial w\right]_m = \partial l(w)/\partial w_m \quad \text{by part (b)}$$

$$= -\sum_n \left[t^{(n)} - w^T x^{(n)}\right] x^{(n)}_m \quad \text{by the note in part (a)}$$

$$= -\sum_n \left[t^{(n)} - \left[Xw\right]_n\right] X_{nm} \quad \text{by part (a)}$$

$$= -\sum_n \left[T_n - \left[Xw\right]_n\right] X_{nm} \quad \text{by the definition of } T$$

$$= -\sum_n \left[T - Xw\right]_n X_{nm}$$

$$= -\sum_n \left[T - Xw\right]_n \left[X^T\right]_{nm} \quad \text{by the definition of matrix transpose}$$

$$= \sum_n \left[X^T\right]_{mn} \left[Xw - T\right]_n \quad \text{by rearranging terms}$$

$$= \left[X^T(Xw - t)\right]_m \quad \text{by the definition of matrix-vector multiplication.}$$

Thus, $\partial l(w)/\partial w = X^T(Xw - t)$ since all their components are equal.