Name: **Student ID:**

CS 189: Introduction to Machine Learning

Homework 1

Due: September 13, 2016 at 11:59pm

Instructions

- This homework includes both a written portion and a coding portion.
- We prefer that you typeset your answers using LaTeX. Neatly handwritten and scanned solutions will also be accepted. Make sure to start each question on a new page.
- You will be submitting **two** things to Gradescope:
  - Append a screenshot or LaTeX snippet of your code to the last page of your writeup. Submit **a PDF of your writeup** to the Homework 1 assignment on Gradescope.
  - Zip up your source code and submit that **zip file** to the Homework 1 Code assignment on Gradescope.
- You should be able to see CS 189/289A on Gradescope when you log in with your bCourses email address. Please make a Piazza post if you have any problems accessing Gradescope.
- The assignment covers concepts in probability, linear algebra, matrix calculus, and decision theory.
- **Start early. This is a long assignment. Some of the material may not have been covered in lecture; you are responsible for finding resources to understand it.**

Problem 1: Expected Value.

A target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Shots within the inner circle are given 4 points, shots within the next ring are given 3 points, and shots within the third ring are given 2 points. Shots outside the target are given 0 points.

Let X be the distance of the hit from the center (in feet), and let the probability density function of X be

$$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value of the score of a single shot? Tip: integration is hard, use Wolfram Alpha.

Solution:

Problem 2: MLE.

Assume that the random variable X has the exponential distribution

$$f(x;\theta) = \theta e^{-\theta x} \qquad x \geq 0, \theta > 0$$

where θ is the parameter of the distribution. Show how to use the method of maximum likelihood to estimate θ from n observations of X: $x_1, \ldots, x_n$.

Solution:

Definition. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. We say that A is **positive definite** if $\forall x \in \mathbb{R}^n \mid x \neq \vec{0},\ x^\top A x > 0$. Similarly, we say that A is **positive semidefinite** if $\forall x \in \mathbb{R}^n,\ x^\top A x \geq 0$.

Problem 3: Positive Definiteness.

Let $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$, and let $A \in \mathbb{R}^{n \times n}$ be the square matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

(a) Give an explicit formula for $x^\top A x$. Write your answer as a sum involving the elements of A and x.

(b) Show that if A is positive definite, then the entries on the diagonal of A are positive (that is, $a_{ii} > 0$ for all $1 \leq i \leq n$).

Solution:

Problem 4: Short Proofs.

A is symmetric in all parts.

(a) Let A be a positive semidefinite matrix. Show that $A + \gamma I$ is positive definite for any $\gamma > 0$.

(b) Let A be a positive definite matrix. Prove that all eigenvalues of A are greater than zero.

(c) Let A be a positive definite matrix. Prove that A is invertible. (Hint: Use the previous part.)

(d) Let A be a positive definite matrix. Prove that there exist n linearly independent vectors $x_1, x_2, ..., x_n$ such that $A_{ij} = x_i^\top x_j$. (Hint: Use the spectral theorem and what you proved in (b) to find a matrix B such that $A = B^\top B$.)

Solution:

Problem 5: Derivatives and Norm Inequalities.

Derive the expression for following questions. Do not write the answers directly.

(a) Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$. Compute $\frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial \mathbf{x}}$.

(b) Let $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{x} \in \mathbb{R}^n$. Compute $\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}}$.

(c) Let $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{n\times n}$. Compute $\frac{\partial \text{Trace}(\mathbf{X}\mathbf{A})}{\partial \mathbf{X}}$.

(d) Let $\mathbf{x} \in \mathbb{R}^n$. Prove that $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2$. (Hint: The Cauchy-Schwarz inequality may come in handy.)

(e) Write down a simple expression for $g(x) = \sup_{\|z\|_1 \leq 1} x^T z$. Hint: first prove an upper bound on $g(x)$, then propose a choice of z that achieves the bound.

Solution:

Problem 6: Gaussian classification.

Let $P(x \mid \omega_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-category, one-dimensional classification problem with $P(\omega_1) = P(\omega_2) = 1/2$. Here, the classes are ω_1 and ω_2. For this problem, we have $\mu_2 \geq \mu_1$.

(a) Find the optimal Bayes decision boundary (i.e., find x such that $P(\omega_1 \mid x) = P(\omega_2 \mid x)$). What is the corresponding decision rule?

(b) Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-z^2/2} dz$$

where $a = \dfrac{\mu_2 - \mu_1}{2\sigma}$. The Bayes error is the probability of misclassification:

$$P_e = P((\text{misclassified as } \omega_1) \mid \omega_2) P(\omega_2) + P((\text{misclassified as } \omega_2) \mid \omega_1) P(\omega_1).$$

Solution:

Problem 7: Regularized Least Squares.

In this question we'll revisit regularized least squares. Let $x_1, \ldots, x_n \in \mathbf{R}^d$, $y_1, \ldots, y_n \in \mathbf{R}$ be the training dataset. Let $X \in \mathbf{R}^{(n,d)}$ be the corresponding data matrix. The ℓ_2-regularized least square estimate for w is the solution to the following optimization problem:

$$\underset{w}{\text{minimize}} \frac{1}{2}\|Xw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \tag{1}$$

Here $\lambda > 0$ is the regularization parameter.

(a) Compute the gradient of the objective function in (1) with respect to w.

(b) Set the gradient to zero to get a closed-form solution for w.

(c) Recall that any vector $w \in \mathbf{R}^d$ can be written as $w = w_n + X^T\alpha$ for some w_n in the nullspace of X (i.e. $Xw_n = 0$) and some $\alpha \in \mathbf{R}^n$. Furthermore, recall that w_n is perpendicular to $X^T\alpha$ for any α. Using this decomposition of w, show that the first term in the objective function of (1) depends only on α, and does not depend on w_n.

(d) Prove that the second term of (1) *does* depend on w_n, but is minimized (over w_n) when $w_n = 0$. Hint: remember that w_n is orthogonal to $X^T\alpha$.

(e) Conclude that $w_\star = X^T\alpha_\star$ for some $\alpha_\star$, and rewrite (1) as an optimization problem over α.

(f) Write down a simple, closed-form solution for $\alpha_\star$. Try to make this as simple as possible.

(g) Compare (f) and (b); computationally, when might you want to find $\alpha_\star$ instead of $w_\star$?

Solution:

Problem 8: Least Squares Classification. In this problem we will implement a least squares classifier for the MNIST data set. The task is to classify handwritten images of numbers between 0 to 9.

We highly recommend you use the anaconda build of python. First you will need to install some packages and get some data.

```
bash code/get_data.sh
pip install python-mnist
pip install sklearn
pip install scipy
pip install numpy
```

Look in `hw1.py` for the skeleton code. You are **NOT** allowed to use any of the prebuilt classifiers in `sklearn`. Feel free to use any method from `numpy` or `scipy`.

a) In this problem we will choose a linear classifier to minimize the least squares objective:

$$W^* = \text{argmin}_{W \in R^{d \times k}} \sum_{i=0}^{n} \|W^T x_i - y_i\|_2^2 + \lambda \|W\|_F^2$$

We adopt the notation where we have n data points and each data point lives in d-dimensional space. k denotes the number of classes. Note that $\|W\|_F$ corresponds to the Frobenius norm of W, i.e. $\|\text{vec}(W)\|_2^2$.

Derive a closed form for $W_\star$.

b) As as first step we need to choose the vectors $y_i \in \mathbf{R}^k$ by converting the original labels (which are in $\{0, \ldots, 9\}$) to vectors.

We will use the one-hot encoding of the labels, i.e. the original label $j \in \{0, \ldots, 9\}$ is mapped to the standard basis vector e_j.

Fill in the function, `one_hot`, that takes a number in $0, \ldots, 9$ and returns the encoded vector.

c) Please implement the functions `train` and `predict` to achieve a test accuracy of 0.85 (that is your classifier should classify 85% of the examples correctly).

The solution to this part should be **very** simple. We have provided the diffstat for the staff solution (this includes all imports). Our solution takes 7 lines of code.

```
hw1.py |   14 +++++++-------
 1 file changed, 7 insertions(+), 7 deletions(-)
```

d) What is the algorithmic run time for computing `train`? Write your answer in $\mathcal{O}$ notation, (In terms of k, d, and n)

e) What could you do speed up training when $n << d$?