STANFORD UNIVERSITY
CS 229, Autumn 2014
Midterm Examination

Wednesday, November 5, 6:00pm-9:00pm

Question	Points
1 Least Squares	/16
2 Generative Learning	/14
3 Generalized Linear Models	/18
4 Support Vector Regression	/18
5 Learning Theory	/20
6 Short Answers	/24
Total	/110

Name of Student: _____

SUNetID: _____@stanford.edu

The Stanford University Honor Code:

I attest that I have not given or received aid in this examination, and that I have done my share and taken an active part in seeing to it that others as well as myself uphold the spirit and letter of the Honor Code.

Signed: _____

1. **[16 points] Least Squares**

   As described in class, in least squares regression we have a cost function:

   $$J(\theta) = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 = (X\theta - \vec{y})^T (X\theta - \vec{y})$$

   The goal of least squares regression is to find θ such that we minimize $J(\theta)$ given the training data.

   Let's say that we had an original set of n features, so that the training inputs were represented by the design matrix $X \in \mathbb{R}^{m \times (n+1)}$. However, we now gain access to one additional feature for every example. As a result, we now have an additional vector of features $\vec{v} \in \mathbb{R}^{m \times 1}$ for our training set that we wish to include in our regression. We can do this by creating a new design matrix: $\widetilde{X} = [X\ \vec{v}] \in \mathbb{R}^{m \times (n+2)}$.

   Therefore the new parameter vector is $\theta_{new} = \begin{pmatrix} \theta \\ p \end{pmatrix}$ where $p \in \mathbb{R}$ is the parameter corresponding to the new feature vector $\vec{v}$.

   Note: For mathematical simplicity, throughout this problem you can assume that $X^T X = I \in \mathbb{R}^{(n+1)\times(n+1)}$ and $\widetilde{X}^T \widetilde{X} = I \in \mathbb{R}^{(n+2)\times(n+2)}, \vec{v}^T \vec{v} = 1$. This is called an *orthonormality* assumption – specifically, the columns of $\widetilde{X}$ are orthonormal. The conclusions of the problem hold even if we do not make this assumption, but this will make your derivations easier.

   (a) [2 points] Let $\hat{\theta} = \arg\min_\theta J(\theta)$ be the minimizer of the original least squares objective (using the original design matrix X). Using the orthornormality assumption, show that $J(\hat{\theta}) = (XX^T\vec{y} - \vec{y})^T (XX^T\vec{y} - \vec{y})$. I.e., show that this is the value of $\min_\theta J(\theta)$ (the value of the objective at the minimum).

(b) [5 points] Now let $\hat{\theta}_{new}$ be the minimizer for $\widetilde{J}(\theta_{new}) = (\widetilde{X}\theta_{new} - \vec{y})^T(\widetilde{X}\theta_{new} - \vec{y})$. Find the new minimized objective $\widetilde{J}(\hat{\theta}_{new})$ and write this expression in the form: $\widetilde{J}(\hat{\theta}_{new}) = J(\hat{\theta}) + f(X, \vec{v}, \vec{y})$ where $J(\hat{\theta})$ is as derived in part (a) and f is some function of $X, \vec{v}$, and $\vec{y}$.

(c) [6 points] Prove that the optimal objective value does not increase upon adding a feature to the design matrix. That is, show $\widetilde{J}(\hat{\theta}_{new}) \leq J(\hat{\theta})$.

(d) [3 points] Does the above result show that if we keep increasing the number of features, we can always get a model that generalizes better than a model with fewer features? Explain why or why not.

2. **[14 points] Decision Boundaries for Generative Models**

   (a) [7 points] Consider the *multinomial event model* of Naive Bayes. Our goal in this problem is to show that this is a linear classifier.

   For a given text document x, let $c_1, ..., c_V$ indicate the number of times each word (out of V words) appears in the document. Thus, $c_i \in \{0, 1, 2, \ldots\}$ counts the occurrences of word i. Recall that the Naive Bayes model uses parameters $\phi_y = p(y = 1)$, $\phi_{i|y=1} = p(\text{word i appears in a specific document position} \mid y = 1)$ and $\phi_{i|y=0} = p(\text{word i appears in a specific document position} \mid y = 0)$.

   We say a classifier is linear if it assigns a label $y = 1$ using a decision rule of the form

$$\sum_{i=1}^{V} w_i c_i + b \geq 0$$

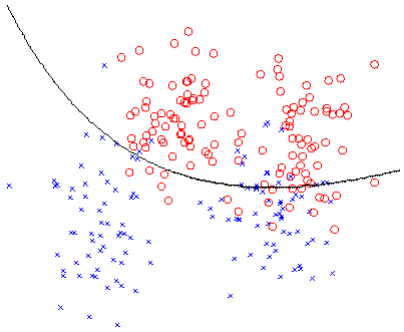
   I.e., the classifier predicts "$y = 1$" if $\sum_{i=1}^{V} w_i c_i + b \geq 0$, and predicts "$y = 0$" otherwise.

   Show that Naive Bayes is a linear classifier, and clearly state the values of w_i and b in terms of the Naive Bayes parameters. (Don't worry about whether the decision rule uses "$\geq$" or "$>$.") Hint: consider using log-probabilities.

[extra space for 2 (a)]

(b) [7 points] In Problem Set 1, you showed that Gaussian Discriminant Analysis (GDA) is a linear classifier. In this problem, we will show that a modified version of GDA has a quadratic decision boundary.

Recall that GDA models $p(x|y)$ using a multivariate normal distribution, where $(x|y=0) \sim \mathcal{N}(\mu_0, \Sigma)$ and $(x|y=1) \sim \mathcal{N}(\mu_1, \Sigma)$, where we used the same Σ for both Gaussians. For this question, we will instead use two covariance matrices Σ_0, Σ_1 for the two labels. So, $(x|y=0) \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $(x|y=1) \sim \mathcal{N}(\mu_1, \Sigma_1)$.



The model distributions can now be written as:

$$
\begin{aligned}
p(y) &= \phi^y (1-\phi)^{1-y} \\
p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma_0|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0)\Big) \\
p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma_1|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\Big)
\end{aligned}
$$

Let's follow a binary decision rule, where we predict $y=1$ if $p(y=1|x) \geq p(y=0|x)$, and $y=0$ otherwise. Show that if $\Sigma_0 \neq \Sigma_1$, then the separating boundary is quadratic in x.

That is, simplify the decision rule "$p(y=1|x) \geq p(y=0|x)$" to the form "$x^T A x + B^T x + C \geq 0$" (supposing that $x \in \mathbb{R}^{n+1}$), for some $A \in \mathbb{R}^{(n+1)\times(n+1)}$, $B \in \mathbb{R}^{n+1}$, $C \in \mathbb{R}$ and $A \neq 0$. Please clearly state your values for A, B and C.

[extra space for 2 (b)]

3. **[18 points] Generalized Linear Models**
In this problem you will build a Generalized Linear Model (GLM) for a response variable y (taking on values $\{1,2,3,\dots\}$), whose distribution (parameterized by ϕ) is modeled as:

$$p(y;\phi) = (1-\phi)^{y-1}\phi$$

This distribution is known as the *geometric distribution*, and is used to model network connections and many other problems.

(a) i. [5 points] Show that the geometric distribution is an exponential family distribution. You should explicitly specify $b(y)$, η, $T(y)$, $\alpha(\eta)$. Also specify what ϕ is in terms of η.

ii. [5 points] Suppose that we have an IID training set $\{(x^{(i)}, y^{(i)}), i = 1, ..., m\}$ and we wish to model this using a GLM based on a geometric distribution. Find the log-likelihood $\log \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}; \theta)$ defined with respect to the entire training set.

(b) [6 points] Derive the Hessian H and the gradient vector of the log likelihood with respect to θ, and state what one step of Newton's method for maximizing the log likelihood would be.

(c) [2 points] Show that the Hessian is negative semi-definite. This shows the optimization objective is concave, and hence Newton's method is maximizing log-likelihood.

4. **[18 points] Support Vector Regression**

   In class, we showed how the SVM can be used for classification. In this problem, we will develop a modified algorithm, called the Support Vector Regression algorithm, which can instead be used for regression, with continuous valued labels $y \in \mathbb{R}$.

   Suppose we are given a training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, where $x^{(i)} \in \mathbb{R}^{(n+1)}$ and $y^{(i)} \in \mathbb{R}$. We would like to find a hypothesis of the form $h_{w,b}(x) = w^T x + b$ with a small value of w. Our (convex) optimization problem is:

$$\begin{array}{lll} \min_{w,b} & \frac{1}{2}\|w\|^2 & \\ \text{s.t.} & y^{(i)} - w^T x^{(i)} - b \leq \epsilon \qquad & i = 1, \ldots, m \quad (1) \\ & w^T x^{(i)} + b - y^{(i)} \leq \epsilon \qquad & i = 1, \ldots, m \quad (2) \end{array}$$

   where $\epsilon > 0$ is a given, fixed value. Notice how the original functional margin constraint has been modified to now represent the distance between the continuous y and our hypothesis' output.

   (a) [4 points] Write down the Lagrangian for the optimization problem above. We suggest you use two sets of Lagrange multipliers α_i and α_i^*, corresponding to the two inequality constraints (labeled (1) and (2) above), so that the Lagrangian would be written $\mathcal{L}(w, b, \alpha, \alpha^*)$.

(b) [10 points] Derive the dual optimization problem. You will have to take derivatives of the Lagrangian with respect to w and b.

[extra space for problem 4 (b)]

(c) [4 points] Show that this algorithm can be kernelized. For this, you have to show that (i) the dual optimization objective can be written in terms of inner-products of training examples; and (ii) at test time, given a new x the hypothesis $h_{w,b}(x)$ can also be computed in terms of inner products.

5. **[20 points] Learning Theory**

Suppose you are given a hypothesis $h_0 \in \mathcal{H}$, and your goal is to determine whether h_0 has generalization error within $\eta > 0$ of the best hypothesis, $h^* = \arg\min_{h \in \mathcal{H}} \varepsilon(h)$. More specifically, we say that a hypothesis h is η**-optimal** if $\varepsilon(h) \leq \varepsilon(h^*) + \eta$. Here, we wish to answer the following question:

Given a hypothesis h_0, is h_0 η-optimal?

Let $\delta > 0$ be some fixed constant, and consider a finite hypothesis class $\mathcal{H}$ of size $|\mathcal{H}| = k$. For each $h \in \mathcal{H}$, let $\hat{\varepsilon}(h)$ denote the training error of h with respect to some training set of m IID examples, and let $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$ denote the hypothesis that minimizes training error.

Now, consider the following algorithm:

1. Set
$$\gamma := \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}.$$
2. If $\hat{\varepsilon}(h_0) > \hat{\varepsilon}(\hat{h}) + \eta + 2\gamma$, then return NO.
3. If $\hat{\varepsilon}(h_0) < \hat{\varepsilon}(\hat{h}) + \eta - 2\gamma$, then return YES.
4. Otherwise, return UNSURE.

Intuitively, the algorithm works by comparing the training error of h_0 to the training error of the hypothesis $\hat{h}$ with the minimum training error, and returns NO or YES only when $\hat{\varepsilon}(h_0)$ is either significantly larger than or significantly smaller than $\hat{\varepsilon}(\hat{h})+\eta$.

(a) [6 points] First, show that if $\varepsilon(h_0) \leq \varepsilon(h^*) + \eta$ (i.e., h_0 is η-optimal), then the probability that the algorithm returns NO is at most δ.

[extra space for 5 (a)]

(b) [6 points] Second, show that if $\varepsilon(h_0) > \varepsilon(h^*) + \eta$ (i.e., h_0 is not η-optimal), then the probability that the algorithm returns YES is at most δ.

(c) [8 points] Finally, suppose that $h_0 = h^*$, and let $\eta > 0$ and $\delta > 0$ be fixed. Show that if m is sufficiently large, then the probability that the algorithm returns YES is at least $1 - \delta$.

Hint: observe that for fixed η and δ, as $m \to \infty$, we have

$$\gamma = \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}} \to 0.$$

This means that there are values of m for which $2\gamma < \eta - 2\gamma$.

6. **[24 points] Short answers**

   The following questions require a reasonably short answer (usually at most 2-3 sentences or a figure, though some questions may require longer or shorter explanations).

   To discourage random guessing, one point will be deducted for a wrong answer on true/false or multiple choice questions! Also, no credit will be given for answers without a correct explanation.

   (a) [3 points] You have an implementation of Newton's method and gradient descent. Suppose that one iteration of Newton's method takes twice as long as one iteration of gradient descent. Then, this implies that gradient descent will converge to the optimal objective faster. True/False?

   (b) [3 points] A stochastic gradient descent algorithm for training logistic regression with a fixed learning rate will always converge to exactly the optimal setting of the parameters $\theta^* = \arg\max_\theta \prod_{i=1}^m p(y^{(i)}|x^{(i)};\theta)$, assuming a reasonable choice of the learning rate. True/False?

(c) [3 points] Given a valid kernel $K(x, y)$ over $\mathbb{R}^m$, is $K_{norm}(x, y) = \frac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}$ a valid kernel?

(d) [3 points] Consider a 2 class classification problem with a dataset of inputs $\{x^{(1)} = (-1, -1), x^{(2)} = (-1, +1), x^{(3)} = (+1, -1), x^{(4)} = (+1, +1)\}$. Can a linear SVM (with no kernel trick) shatter this set of 4 points?

(e) [3 points] For linear hypotheses (i.e. of the form $h(x) = w^T x + b$), the vector of learned weights w is always perpendicular to the separating hyperplane. True/False? Provide a counterexample if False, or a brief explanation if True.

(f) [3 points] Let $\mathcal{H}$ be a set of classifiers with a VC dimension of 5. Consider a set of 5 training examples $\{(x^{(1)}, y^{(1)}), ..., (x^{(5)}, y^{(5)})\}$. Now we select a classifier h^* from $\mathcal{H}$ by minimizing the classification error on the training set. Which one of the following is true?

i. $x^{(5)}$ will certainly be classified correctly (i.e. $h^*(x^{(5)}) = y^{(5)}$)

ii. $x^{(5)}$ will certainly be classified incorrectly (i.e. $h^*(x^{(5)}) \neq y^{(5)}$)

iii. We cannot tell

Briefly justify your answer.

(g) [6 points] Suppose you would like to use a linear regression model in order to predict the price of houses. In your model, you use the features $x_0 = 1$, $x_1 =$ size in square meters, $x_2 =$ height of roof in meters. Now, suppose a friend repeats the same analysis using exactly the same training set, only he represents the data instead using features $x'_0 = 1$, $x'_1 = x_1$, and $x'_2 =$ height in cm (so $x'_2 = 100x_2$).

i. [3 points] Suppose both of you run linear regression, solving for the parameters via the Normal equations. (Assume there are no degeneracies, so this gives a unique solution to the parameters.) You get parameters θ_0, θ_1, θ_2; your friend gets θ'_0, θ'_1, θ'_2. Then $\theta'_0 = \theta_0$, $\theta'_1 = \theta_1$, $\theta'_2 = \frac{1}{100}\theta_2$. True/False?

ii. [3 points] Suppose both of you run linear regression, initializing the parameters to 0, and compare your results after running just *one* iteration of batch gradient descent. You get parameters θ_0, θ_1, θ_2; your friend gets θ'_0, θ'_1, θ'_2. Then $\theta'_0 = \theta_0$, $\theta'_1 = \theta_1$, $\theta'_2 = \frac{1}{100}\theta_2$. True/False?