STANFORD UNIVERSITY
CS 229, Autumn 2016
Midterm Examination

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

Question	Points
1 Short answers	/24
2 Linear regression	/12
3 Generative models	/12
4 Generalized linear models	/22
5 Kernels	/16
6 Learning theory	/10
Total	/96

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. **[24 points] Short answers**

   The following questions require a reasonably short answer (usually at most 2-3 sentences or a figure for each question part, though some 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) [5 points] Given a cost function $J(\theta)$ that we seek to minimize and $\alpha \in \mathbb{R} > 0$, consider the following update rule:

   $$\theta^{(t+1)} = \arg\min_{\theta} \left\{ J(\theta^{(t)}) + \nabla_{\theta^{(t)}} J(\theta^{(t)})^T (\theta - \theta^{(t)}) + \frac{1}{2\alpha} \|\theta - \theta^{(t)}\|_2^2 \right\}.$$

   i. [3 points] Show that this yields the same $\theta^{(t+1)}$ as the gradient descent update with step size α.

   ii. [2 points] Provide a sketch (i.e. draw a picture) of the above update for the simplified case where $\theta \in \mathbb{R}$, $J(\theta) = \theta$, and $\theta^{(t)} = 1$. Make sure to clearly label $\theta^{(t)}$, $\theta^{(t+1)}$ and α.

(b) [4 points] In the binary classification setting where $y \in \{-1, +1\}$, we define the margin as $z = y\theta^T x$ where θ and x lie in $\mathbb{R}^n$. Consider each of the following three loss functions:

i. zero-one loss: $\varphi_{\text{zo}}(z) = 1\{z \leq 0\}$ i cuz gradient = 0

ii. exponential loss: $\varphi_{\text{exp}}(z) = e^{-z}$

iii. hinge loss: $\varphi_{\text{hinge}}(z) = \max\{1 - z, 0\}$

Suppose we have margin $z < 0$ for our current parameters θ. Give the expression for $\frac{\partial}{\partial \theta_k} \varphi(y\theta^T x)$ for each of the given loss functions. Which loss would we fail to minimize with gradient descent, no matter the step size we choose?

(c) [5 points] Consider performing spam classification where each e-mail is represented as a vector x of the same size as the number of words in the vocabulary $|V|$, where x_i is 1 if the the e-mail contains word i and 0 otherwise. We saw in class that Naive Bayes with Laplace smoothing is one simple method for performing classification in this setting. For this question, to simplify we set $p(y = 1) = p(y = -1) = 0.5$.

Consider classifying x by instead using the boosting algorithm with $2|V|$ decision stumps as the weak learners. In this setting, which of the two methods, Naive Bayes or boosting with decision stumps, would you expect to yield lower bias? Explain your reasoning. naive bayes since it assumes the independence between features

(d) [4 points] Suppose we trained a linear SVM classifier to perform binary classification using the hinge loss $L(\theta^T x, y) = \max\{0, 1 - y\theta^T x\}$. For each of the following scenarios, does the optimal decision boundary necessarily remain the same? Explain your reasoning, perhaps by sketching a picture. Assume that after we perform the action described in each scenario we still have at least one training example in the positive class as well as in the negative class.

i. Remove all examples $(x^{(i)}, y^{(i)})$ with margin > 1.

ii. Remove all examples $(x^{(i)}, y^{(i)})$ with margin < 1.

iii. Add an ℓ_2-regularization term $\frac{\lambda}{2}\theta^T\theta = \frac{\lambda}{2}\|\theta\|_2^2$ to the training loss.

iv. Scale all $x^{(i)}$ by a constant factor α.

(e) [6 points] We consider a binary classification task where we have m training examples and our hypothesis $h_\theta(x)$ is parameterized by θ. For each of the following scenarios, select whether we should expect bias and variance to increase or decrease. Explain your reasoning.

i. Project the values of θ to lie between -1 and 1 after each training update, that is $\theta_j = \min\{1, \max\{-1, \theta_j\}\}$.

theta could get only 2 values (-1 and 1) => decrease hypothesis => decrease bias but increase variance space

bias: increase decrease **variance:** increase decrease

ii. Smooth the estimates of our hypotheses by outputting

$$h(x) = (1/3) \sum_{x^{(i)} \in N_3(x)} h_\theta(x^{(i)}),$$

encourage similar outputs => same result

where $N_3(x)$ are the 3 points in the training set closest to x.

bias: increase decrease **variance:** increase decrease

iii. Remove one of the feature dimensions of x.

the hypothesis space is now a strict subset of preivous

bias: increase decrease **variance:** increase decrease

2. **[12 points] Linear regression: First order convergence for least squares**

   Consider the least squares problem, where we pick θ to minimize the objective $J(\theta) = \frac{1}{2}(X^T\theta - y)^T(X^T\theta - y)$. The solution to this problem is given by the normal equation, where $\theta = (XX^T)^{-1}Xy$. In Problem Set 1, we showed that a single Newton step will converge to the correct solution. Now we will examine how gradient descent performs on the same problem.

   (a) [4 points] Find the gradient of J with respect to θ, and write the gradient descent update step for $\theta^{(t+1)}$, given $\theta^{(t)}$ and step size α.

(b) [8 points] Show that as $t \to \infty$, $\theta^{(t+1)} \to (XX^T)^{-1}Xy$, for gradient descent with step size α and $\theta^{(0)} = 0$. You may use the fact that $(\alpha A)^{-1} = \sum_{i=0}^{\infty} (I - \alpha A)^i$ for small $\alpha > 0$, and you may assume that your choice of α is small enough.

3. **[12 points] Generative models: Gaussian discriminant analysis, continued**

   Consider the 1-dimensional Gaussian discriminant analysis model where $x \in \mathbb{R}$ and we assume

$$\begin{aligned} p(y) &= \phi^{1\{y=1\}}(1-\phi)^{1\{y=-1\}} \\ p(x|y=-1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_{-1})^2\right) \\ p(x|y=1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_{1})^2\right) \end{aligned}$$

   In this problem we will assume that σ *is a fixed quantity* that we have been given and is therefore not a parameter of the model.

   Recall from Problem Set 1 that we can express $p(y|x;\phi,\mu_{-1},\mu_1)$ in the form

$$p(y|x;\theta) = \frac{1}{1+\exp(-y(\theta_1 x + \theta_0))}$$

   where for the model described above we have,

$$\begin{aligned} \theta_0 &= \frac{1}{2\sigma^2}(\mu_{-1}^2 - \mu_1^2) - \log\frac{1-\phi}{\phi} \\ \theta_1 &= \frac{1}{\sigma^2}(\mu_1 - \mu_{-1}). \end{aligned}$$

   (a) [2 points] Write the joint log-likelihood $\ell(\phi,\mu_{-1},\mu_1) = \log p(x,y;\phi,\mu_{-1},\mu_1)$ for a single example (x,y).

(b) [7 points] Show that the log-likelihood of all training examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$ is concave (and hence any maximum we find must be the global maximum) by first computing $\frac{\partial^2 \ell}{\partial \phi^2}$, $\frac{\partial^2 \ell}{\partial \mu_{-1}^2}$, and $\frac{\partial^2 \ell}{\partial \mu_1^2}$ for a single example (x, y). Then make an argument that the total log-likelihood is concave. Hint: Recall a function is concave if its Hessian is negative semidefinite. A one-dimensional function f is concave if $f''(x) \leq 0$ for all x.

(c) [3 points] Derive an expression for the decision boundary for classifying x as either $y = -1$ or 1.

4. **[22 points] Generalized linear models: Gaussian distribution**

   Assume we are given $x_1, x_2, \ldots, x_n$ drawn i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, that is,

   $$p(x_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)$$

   Define $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$ where $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$.

   (a) [3 points] Prove $g(x) = \frac{s^2}{n-1}$ is an unbiased estimator of σ^2, that is

   $$\mathbb{E}[g(x)] = \sigma^2$$

   Hint: $\mathbb{E}[x_i] = \mu, \mathrm{Var}(x_i) = \sigma^2, \mathrm{Cov}(x_i, x_j) = 0$.

(b) [5 points] Find the maximum-likelihood estimate of μ and σ^2. Hint: You should be able to express your final expression for σ^2 in terms of s^2.

(c) [6 points] Show that the general form of the Gaussian distribution is a member of the exponential family by finding $b(x)$, η, $T(x)$, and $a(\eta)$. Hint: Since both μ and σ^2 are parameters, η and $T(x)$ will now be two dimensional vectors. Denote $\eta = [\eta_1, \eta_2]^T$ and try to express $a(\eta)$ in terms of η_1 and η_2.

(d) [4 points] Verify that $\nabla_\eta a(\eta) = \mathbb{E}[T(x); \eta]$ for the Gaussian distribution. Hint: You can prove this either by using the general form of exponential families, or by computing $\nabla_\eta a(\eta)$ directly from part (c).

(e) [4 points] Show that $\nabla_\eta^2 a(\eta)$ is positive semidefinite. Hint: You can compute $\nabla_\eta^2 a(\eta)$ using the results from part (c) and (d). Or instead you may use the following fact: In general for exponential families,

$$\nabla_\eta^2 a(\eta) = \mathbb{E}\left[T(x)T(x)^T\right] - \mathbb{E}[T(x)]\mathbb{E}[T(x)]^T$$

.

5. **[16 points] Shift Invariant Kernels**

   A kernel K on $\mathbb{R}^n$ is said to be shift invariant if:

   $$\forall \delta \in \mathbb{R}^n, \forall x, z \in \mathbb{R}^n, K(x, z) = K(x + \delta, z + \delta)$$

   (a) [4 points] Give an example of a shift invariant and a non-shift invariant kernels seen in lectures (no need to prove they are kernels). For the rest of this problem, we will simplify a bit and consider the case where $n = 1$.

(b) [6 points] Let $p(\omega)$ be a probability density over $\mathbb{R}$ and ϕ be a function mapping $\mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^d$. We define $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as:

$$F(x, z) = \int_{-\infty}^{\infty} \phi(x, \omega)^T \phi(z, \omega) p(\omega) \mathrm{d}\omega$$

for all $x, z \in \mathbb{R}^n$. Show that F is a kernel.

(c) [4 points] Let's suppose $n = 1$. Let $h : \mathbb{R} \to \mathbb{R}$ be a function such that

$$\forall z \in \mathbb{R}, h(z) = \int_{-\infty}^{\infty} \cos(\omega z) p(\omega) \mathrm{d}\omega$$

Show that there exists ϕ such that $h(x - z) = \int_{-\infty}^{\infty} \phi(x, \omega)^T \phi(z, \omega) p(\omega) \mathrm{d}\omega$. Provide an explicit definition of ϕ. Hint: Use the trigonometric identity $\cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b)$, valid for all $a, b \in \mathbb{R}$.

(d) [2 points] Show that $K(x, z) = h(x - z)$ is indeed a kernel.

6. **[10 points] Learning theory: Relaxed generalization bounds**

Let $Z_1, Z_2, \ldots, Z_m$ be independent and identically distributed random variables drawn from a Bernoulli(ϕ) distribution where $P(Z_i = 1) = \phi$ and $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} = (1/m) \sum_{i=1}^{m} Z_i$, and let any $\gamma > 0$ be fixed. Hoeffding's inequality, as we saw in class, states

$$\mathbb{P}(|\phi - \hat{\phi}| > \gamma) \leq 2 \exp(-2\gamma^2 m)$$

However, this relies on the assumption that the random variables $Z_1, \ldots, Z_m$ are all *jointly* independent. In this problem we will relax this assumption by only assuming *pairwise* independence among the Z_i. In this case we cannot apply Hoeffding's inequality, but the following inequality (Chebyshev's inequality) holds:

$$P(|\phi - \hat{\phi}| > \gamma) \leq \frac{\text{Var}(Z_i)}{m\gamma^2}$$

where $\text{Var}(Z_i)$ denotes the variance of the random variable Z_i and for $Z_i \sim \text{Bernoulli}(\phi)$ we have $\text{Var}(Z_i) = \phi(1 - \phi)$.

Given our hypothesis set $\mathcal{H} = \{h_1, \ldots, h_k\}$ and m pairwise but not necessarily jointly independent data samples $(x, y) \sim \mathcal{D}$, we now derive guarantees on the generalization error of our best hypothesis

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{argmin}}\, \hat{\varepsilon}(h)$$

where as usual we define $\hat{\varepsilon}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x^{(i)}) \neq y^{(i)}\}$, where $(x^{(i)}, y^{(i)})$ are examples from the training set.

(a) [2 points] What is the maximum possible value of $\text{Var}(Z_i) = \phi(1 - \phi)$? From now on we will instead use this maximal value such that the bounds we derive hold for all possible ϕ.

(b) [4 points] Let $\gamma > 0$.

i. [2 points] Give a non-trivial (i.e. not the constant 1) upper bound on the probability that $|\hat{\varepsilon}(\hat{h}) - \varepsilon(\hat{h})| > \gamma$.

ii. [1 points] Fix $\delta \in (0, 1)$. Using your upper bound, how large must the sample size m be before you can guarantee that

$$\mathbb{P}(|\hat{\varepsilon}(\hat{h}) - \varepsilon(\hat{h})| > \gamma) \leq \delta,$$

that is, that the training error and generalization error are within γ of one another with probability at least $1 - \delta$?

iii. [1 points] How does this sample size compare to what is achievable using Hoeffding's inequality?

(c) [4 points] Show that with probability at least $1 - \delta$, the difference between the generalization error of $\hat{h}$ and the generalization error of the best hypothesis in $\mathcal{H}$ (i.e. the hypothesis $h^\star = \text{argmin}_{h \in \mathcal{H}} \varepsilon(h)$) is bounded by $\sqrt{k/(m\delta)}$.