# STANFORD UNIVERSITY
# CS 229, Autumn 2015
# Midterm Examination

Wednesday, November 4, 6:00pm-9:00pm

| Question | Points |
|---|---|
| 1 Short Answers | /26 |
| 2 More Linear Regression | /10 |
| 3 Generalized Linear Models | /17 |
| 4 Naive Bayes and Logistic Regression | /17 |
| 5 Anomaly Detection | /15 |
| 6 Learning Theory | /15 |
| Total | /100 |

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. **[26 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) [6 points] Suppose you are fitting a fixed dataset with $m$ training examples using linear regression, $h_\theta(x) = \theta^T x$, where $\theta, x \in \mathbb{R}^{n+1}$. After training, you realize that the variance of your model is relatively high (i.e. you are overfitting). For the following methods, indicate true if the method can mitigate your overfitting problem and false otherwise. Briefly explain why.

i. [3 points] Add additional features to your feature vector.

ii. [3 points] Impose a prior distribution on $\theta$, where the distribution of $\theta$ is of the form $\mathcal{N}(0, \tau^2 I)$, and we derive $\theta$ via maximum a posteriori estimation.

(b) [3 points] Choosing the parameter $C$ is often a challenge when using SVMs. Suppose we choose $C$ as follows: First, train a model for a wide range of values of $C$. Then, evaluate each model on the test set. Choose the $C$ whose model has the best performance on the test set. Is the performance of the chosen model on the test set a good estimate of the model's generalization error?

(c) [11 points] For the following, provide the VC-dimension of the described hypothesis classes and briefly explain your answer.

i. [3 points] Assume $\mathcal{X} = \mathbb{R}^2$. $\mathcal{H}$ is a hypothesis class containing a single hypothesis $h_1$ (i.e. $\mathcal{H} = \{h_1\}$)

ii. [4 points] Assume $\mathcal{X} = \mathbb{R}^2$. Consider $\mathcal{A}$ to be the set of all convex polygons in $\mathcal{X}$. $\mathcal{H}$ is the class of all hypotheses $h_P(x)$ (for $P \in \mathcal{A}$) such that

$$h_P(x) = \begin{cases} 1 & \text{if } x \text{ is contained within polygon P} \\ 0 & \text{otherwise} \end{cases}$$

*Hint: Points on the edges or vertices of $P$ are included in $P$*

iii. [4 points] $\mathcal{H}$ is the class of hypotheses $h_{(a,b)}(x)$ such that each hypothesis is represented by a single open interval in $\mathcal{X} = \mathbb{R}$ as follows:

$$h_{(a,b)}(x) = \begin{cases} 1 & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

(d) [3 points] Consider a sine function $f(x) = \sin(x)$ such that $x \in [-\pi, \pi]$. We use two different hypothesis classes such that $\mathcal{H}_0$ contains all constant hypotheses of the form, $h(x) = b$ and $\mathcal{H}_1$ contains all linear hypotheses of the form $h(x) = ax + b$. Consider taking a very large number of training sets, $S_i, i = 1, ..., N$ such that each $S_i$ contains only two points $\{(x_1, y_1), (x_2, y_2)\}$ sampled iid from $f(x)$. In other words, each $(x, y)$ pair is drawn from a distribution such that $y = f(x) = \sin(x)$ is satisfied. We train a model from each hypothesis class using each training set such that we have a collection of $N$ models from each class. We then compute a mean-squared error between each model and the function $f(x)$.

It turns out that the average expected error of all models from $\mathcal{H}_0$ is significantly lower than the average expected error of models from $\mathcal{H}_1$ even though $\mathcal{H}_1$ is a more complex hypothesis class. Using the concepts of bias and variance, provide an explanation for why this is the case.

(e) [3 points] In class when we discussed the decision boundary for logistic regression $h_\theta(x) = g(\theta^T x)$, we did not require an explicit intercept term because we could define $x_0 = 1$ and let $\theta_0$ be the intercept. When discussing SVMs, we dropped this convention and had $h_{w,b}(x) = g(w^T x + b)$ with $b$ as an explicit intercept term. Consider an SVM where we now write $h_w(x) = g(w^T x)$ and define $x_0 = 1$ such that $w_0$ is the intercept. If the primal optimization objective remains $\frac{1}{2} ||w||^2$, can we change the intercept in this way without changing the decision boundary found by the SVM? Justify your answer.

2. **[10 + 3 Extra Credit points] More Linear Regression**

In our homework, we saw a variant of linear regression called locally-weighted linear regression. In the problem below, we consider a regularized form of locally-weighted linear regression where we favor smaller parameter vectors by adding a complexity penalty term to the cost function. Additionally, we consider the case where we are trying to predict multiple outputs for each training example. Our dataset is:

$$\mathcal{S} = \{(x^{(i)}, y^{(i)}), i = 1, ..., m\}, x^{(i)} \in \mathbb{R}^n, y^{(i)} \in \mathbb{R}^p$$

Thus for each training example, $y^{(i)}$ is a real-valued vector with $p$ entries. We wish to use a linear model to predict the outputs by specifying the parameter matrix $\theta$, where $\theta \in \mathbb{R}^{n \times p}$. You can assume $x^{(i)}$ contains the intercept term (i.e. $x_0 = 1$). The cost function for this model is:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{p} w^{(i)} \left( (\theta^T x^{(i)})_j - y_j^{(i)} \right)^2 + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{p} (\theta_{ij})^2 \qquad (1)$$

As before, $w^{(i)}$ is the "weight" for a specific training example $i$.

(a) [2 points] Show that $J(\theta)$ can be written as

$$J(\theta) = \frac{1}{2} \mathrm{tr} \left( (X\theta - Y)^T W (X\theta - Y) \right) + \frac{1}{2} \mathrm{tr}(\theta^T \theta)$$

(b) [5 points] Derive a closed form expression for the minimizer $\theta^*$ that minimizes $J(\theta)$ from part (a).

(c) [3 points] Given the dataset $\mathcal{S}$ above, which of the following cost functions will lead to higher accuracy on the training set? Briefly explain why this is the case. If there is insufficient information, explain what details are needed to make a decision.

i. $J_1(\theta) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}\left((\theta^T x^{(i)})_j - y_j^{(i)}\right)^2$

ii. $J_2(\theta) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}\left((\theta^T x^{(i)})_j - y_j^{(i)}\right)^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}(\theta_{ij})^2$

iii. $J_3(\theta) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}\left((\theta^T x^{(i)})_j - y_j^{(i)}\right)^2 + 100\sum_{i=1}^{n}\sum_{j=1}^{p}(\theta_{ij})^2$

(d) [3 Extra Credit points] Suppose we want to weight the regularization penalty on a per element basis. For this problem, we use the following cost function:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p} w^{(i)} \left((\theta^T x^{(i)})_j - y_j^{(i)}\right)^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}((\Gamma\theta)_{ij})^2 \tag{15}$$

Here, $\Gamma \in \mathbb{R}^{n\times n}$ where $\Gamma_{ij} > 0$ for all $i, j$. Derive a closed form solution for $J(\theta)$ and $\theta^*$ using this new cost function.

3. **[17 points] Generalized Linear Models**

In class we showed that the Gaussian distribution is in the Exponential Family. However, a simplification we made to make the derivation easier was to set the variance term $\sigma^2 = 1$. This problem will investigate a more general form for the Exponential Family. First, recall that the Gaussian distribution can be written as follows:

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \tag{29}$$

(a) [6 points] Show that the Gaussian distribution (without assuming unit variance) is an exponential family distribution. In particular, please specify $b(y)$, $\eta$, $T(y)$, $a(\eta)$. Recall that the standard form for the exponential family is given by

$$p(y;\eta) = b(y)\exp\{\eta^\top T(y) - a(\eta)\} \tag{30}$$

*Hint: since $\sigma^2$ is now a variable, $\eta$ and $T(y)$ will now be two dimensional vectors; for consistent notation denote $\eta = \begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}^\top$. For full credit, please ensure $a(\eta)$ is expressed in terms of $\eta_1$ and $\eta_2$.*

(b) [4 points] Suppose you are given an IID training set $\{(x^{(i)}, y^{(i)}), i = 1, ..., m\}$. Starting with the expression in (30) for $p(y; \eta)$, derive the general expression for the Hessian of the log-likelihood $\ell(\theta) = \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)}; \theta)$. Your answer should be in terms of $x$, $\eta_1$ and $\eta_2$.

(c) [5 points] Using your result from the part (b), show that the Hessian is negative semi-definite, i.e., $z^\top H z \leq 0$.

(d) [2 points] It turns out there is a more general definition for the exponential family given by

$$p(y;\eta,\tau) = b(a,\tau)\exp\left\{\frac{\eta^\top T(y) - a(\eta)}{c(\tau)}\right\}$$

In particular $c(\tau)$ is the dispersion function, where $\tau$ is called the *dispersion parameter*. Show that the Gaussian distribution can be written in this more general form with $c(\tau) = \sigma^2$.

4. **[17 points] Naive Bayes and Logistic Regression**
For this entire problem assume that the input features $x_j$, $j = 1, ..., n$ are discrete binary-valued variables such that $x_j \in \{0, 1\}$ and $x = [x_1 \; x_2 \; ... \; x_n]$. For each training example $x^{(i)}$, assume that the output target variable $y^{(i)} \in \{0, 1\}$.

   (a) [2 points] Consider the Naive Bayes model, given the above context. This model can be parameterized by $\phi_{j|y=0} = p(x_j = 1|y = 0)$, $\phi_{j|y=1} = p(x_j = 1|y = 1)$ and $\phi_y = p(y = 1)$. Write down the expression for $p(y = 1|x)$ in terms of $\phi_{j|y=0}, \phi_{j|y=1}$, and $\phi_y$.

(b) [7 points] Show that the conditional likelihood expression you obtained in part (a) can be simplified to the same form as the hypothesis for logistic regression:

$$p(y = 1|x) = \frac{1}{1 + e^{-\theta^T x}}. \tag{33}$$

*Hint: Modify the definition of $x$ to include the intercept term $x_0 = 1$*

[More space for (b)]

(c) [6 points] In part (b) you showed that the discrete Naive Bayes decision boundary has the same form as that of the logistic regression. Now consider a dataset $S_1$ with $m$ training examples of the form: $\{(x^{(i)}, y^{(i)}), i = 1, ..., m\}$ with each $x^{(i)} \in \mathbb{R}^{n+1}$. Note that for this problem, $S_1$ satisfies the Naive Bayes assumption: $p(x_1, \ldots, x_n|y) = \prod_{j=1}^{n} p(x_j|y)$.
Suppose a second dataset, $S_2$, is given to you again with $m$ training examples $\{(x^{(i)}, y^{(i)}), i = 1, \cdots, m\}$, but now each $x^{(i)} \in \mathbb{R}^{n+2}$ because each $x^{(i)}$ contains the same $n$ conditionally-independent features and an additional feature $x_{n+1}$ such that $x_{n+1} = x_n$. Each $x^{(i)}$ contains the intercept term $x_0 = 1$.

i. [2 points] You train two Naive Bayes classifiers independently on $S_1$ and $S_2$. Test data is generated according to the true distribution (i.e. $p(x_1, ..., x_n, y) = p(x_1, ..., x_n, x_{n+1}, y) = p(y)p(x_1, ..., x_n|y)$, where $x_{n+1} = x_n$). Would you expect the test error of the classifier trained on $S_1$ to be larger or smaller than that trained on $S_2$? You may assume that $m$ is very large. Briefly justify your answer.

ii. [4 points] Now we will look at a similar situation regarding how logistic regression is affected by copies of features. In order to simplify the math, let's assume a more basic case where $S_1$ still has $m$ training examples, but now has one feature $x_1$. $S_2$ has $m$ training examples but has two features $x_1$ and $x_2$ where $x_2 = x_1$. The logistic regression model trained on $S_1$ therefore has associated parameters $\{\theta_0, \theta_1\}$ and the model trained on $S_2$ has parameters $\{\theta_0, \theta_1, \theta_2\}$. Here, $\theta_0$ is associated with the intercept term $x_0 = 1$. Testing data is generated the same way (from the original true distribution). How will the error of the classifier trained on $S_1$ compare to that of the classifier trained on $S_2$? For this question you need to prove your result mathematically. (Hint: compare the forms of the log-likelihood for each classifier)

(d) [2 points] In general, if we assume that the number of training examples $m$ is very large, which classifier will have a lower generalization error? Briefly justify why.

5. **[15 points] Anomaly Detection**

Consider the following optimization problem:

$$\begin{aligned} \underset{r,z,\xi}{\text{minimize}} \quad & r^2 + C\sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & \left|\left|x^{(i)} - z\right|\right|_2^2 \le r^2 + \xi_i \ \ i = 1, \ldots, m. \\ & \xi_i \ge 0, \ \ i = 1, \ldots, m. \end{aligned} \tag{40}$$

where $\xi_i$ are the slack variables.

(a) [2 points] Write down the Lagrangian for the optimization problem above. We suggest using two sets of Lagrange multipliers $\alpha_i$ and $\eta_i$ corresponding to the two inequality constraints so that the Lagrangian would be written as $\mathcal{L}(r, z, \xi, \alpha, \eta)$.

(b) [7 points] Assuming a non-trivial solution ($r > 0$), derive the dual optimization problem using the Lagrangian from part (a).

(c) [3 points] Show that the dual problem from (b) can be kernelized.

(d) [3 points] Now consider the following dual optimization problem

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle \\ \text{s.t} \quad & \sum_{i=1}^{m}\alpha_i = 1, \quad i = 1, \ldots, m. \end{aligned} \tag{49}$$

*Assume that we choose $K$ such that it is a Gaussian Kernel.* How does this dual compare with the dual you derived in part (c)?.

6. **[15 points] Learning Theory**

Consider a finite hypothesis class $\mathcal{H}$ with size $k = |\mathcal{H}|$ and $h^\star = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\ \epsilon(h)$.

(a) [7 points] Assume that the best hypothesis $h^\star$ has generalization error $\epsilon(h^\star) = B$ such that $B$ is a constant with $0 \leq B \leq 1$. Prove that the joint probability of the expected risk minimizer $\hat{h}$ having large generalization error and the best hypothesis $h^*$ having small training error can be bounded as:

$$P(\epsilon(\hat{h}) > B + 2\gamma, \hat{\epsilon}(h^\star) \leq B + \gamma) \leq \sum_{h \in \mathcal{H}} P(\epsilon(h) > B + 2\gamma, \hat{\epsilon}(h) \leq B + \gamma) \quad (50)$$

For any hypothesis $h' \in \mathcal{H}$ with high generalization error (i.e. $\epsilon(h') > B' + \tau$), the probability that it has low training error (i.e. $\hat{\epsilon}(h') \leq B'$) is bounded by:

$$P(\hat{\epsilon}(h') \leq B' \mid \epsilon(h') > B' + \tau) \leq \exp\left\{\frac{-m\tau^2}{2(B' + 4\tau/3)}\right\} \tag{51}$$

for any $B' \in (0, 1)$ and $\tau > 0$.

(b) [8 points] Using (51) and the result from part (a), show that:

$$P(\epsilon(\hat{h}) > B + 2\gamma, \hat{\epsilon}(h^{\star}) \leq B + \gamma) \leq k \exp\left\{\frac{-m\gamma^2}{2(B + 7\gamma/3)}\right\}. \tag{52}$$