CS 189 Spring 2015 Introduction to Machine Learning Midterm

- You have 80 minutes for the exam.
- The exam is closed book, closed notes except your one-page crib sheet.
- No calculators or electronic items.
- Mark your answers ON THE EXAM ITSELF. If you are not sure of your answer you may wish to provide a *brief* explanation.
- For true/false questions, fill in the *True/False* bubble.
- For multiple-choice questions, fill in the bubbles for **ALL** CORRECT CHOICES (in some cases, there may be more than one). We have introduced a negative penalty for false positives for the multiple choice questions such that the expected value of randomly guessing is 0. Don't worry, for this section, your score will be the maximum of your score and 0, thus you cannot incur a negative score for this section.

First name	
Last name	
SID	
First and last name of student to your left	
First and last name of student to your right	

For staff use only:

Q1.	True or False	/26
Q2.	Multiple Choice	/36
Q3.	Parameter Estimation	/10
Q4.	Dual Solution for Ridge Regression	/8
Q5.	Regularization and Priors for Linear Regression	/8
	Total	/88

Q1. [26 pts] True or False

(a) [2 pts] If the data is not linearly separable, then there is no solution to the hard-margin SVM.

● True ○ False

(b) [2 pts] Logistic Regression can be used for classification.

● True ○ False

(c) [2 pts] In logistic regression, two ways to prevent β vectors from getting too large are using a small step size and using a small regularization value.

○ True ● False

(d) [2 pts] The L2 norm is often used because it produces sparse results, as opposed to the L1 norm which does not.

○ True ● False

(e) [2 pts] For a Multivariate Gaussian, the eigenvalues of the covariance matrix are inversely proportional to the lengths of the ellipsoid axes that determine the isocontours of the density.

○ True ● False

(f) [2 pts] In a generative binary classification model where we assume the class conditionals are distributed as Poisson, and the class priors are Bernoulli, the posterior assumes a logistic form.

● True ○ False

(g) [2 pts] Maximum likelihood estimation gives us not only a point estimate, but a distribution over the parameters that we are estimating.

○ True ● False

(h) [2 pts] Penalized maximum likelihood estimators and Bayesian estimators for parameters are better used in the setting of low-dimensional data with many training examples as opposed to the setting of high-dimensional data with few training examples.

○ True ● False

(i) [2 pts] It is not a good machine learning practice to use the test set to help adjust the hyperparameters of your learning algorithm.

● True ○ False

(j) [2 pts] A symmetric positive semi-definite matrix always has nonnegative elements.

○ True ● False

(k) [2 pts] For a valid kernel function K, the corresponding feature mapping ϕ can map a finite dimensional vector into an infinite dimensional vector.

● True ○ False

(l) [2 pts] The more features that we use to represent our data, the better the learning algorithm will generalize to new data points.

○ True ● False

(m) [2 pts] A discriminative classifier explicitly models $P(Y|X)$

● True ○ False

Q2. [36 pts] Multiple Choice

(a) [3 pts] Which of the following algorithms can you use kernels with?

- ● Support Vector Machines
- ● Perceptrons
- ○ None of the above

(b) [3 pts] Cross validation:

- ● Is often used to select hyperparameters
- ○ Is guaranteed to prevent overfitting
- ○ Does nothing to prevent overfitting
- ○ None of the above

(c) [3 pts] In linear regression, L2 regularisation is equivalent to imposing a:

- ○ Logistic prior
- ● Gaussian prior
- ○ Laplace prior
- ○ Gaussian class-conditional

(d) [3 pts] Say we have two 2-dimensional Gaussian distributions representing two different classes. Which of the following conditions will result in a linear decision boundary:

- ○ Same mean for both classes
- ● Same covariance matrix for both classes
- ○ Different covariance matrix for each class
- ○ Linearly separable data

(e) [3 pts] The normal equations can be derived from:

- ● Minimizing empirical risk
- ● Assuming that $Y = \beta^T x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.
- ● Assuming that the $P(Y|X = x)$ is distributed normally with mean $\beta^\intercal x$ and variance σ^2
- ○ Finding a linear combination of the rows of the design matrix that minimizes the distance to our vector of labels Y

(f) [3 pts] Logistic regression can be motivated from:

- ○ Generative models with uniform class conditionals
- ● Generative models with Gaussian class conditionals
- ● Log odds being equated to an affine function of x
- ○ None of the above

(g) [3 pts] The perceptron algorithm will converge:

- ● If the data is linearly separable
- ○ Even if the data is linearly inseparable
- ○ As long as you initialize θ to all 0's
- ○ Always

(h) [3 pts] Which of the following is true:

- [x] Newton's Method typically is more expensive to calculate than gradient descent, per iteration
- [x] For quadratic equations, Newton's Method typically requires fewer iterations than gradient descent
- [ ] Gradient descent can be viewed as iteratively reweighted least squares
- [ ] None of the above

(i) [3 pts] Which of the following statements about duality and SVMs is (are) true?

- [x] Complementary slackness implies that every training point that is misclassified by a soft-margin SVM is a support vector.
- [x] When we solve the SVM with the dual problem, we need only the dot product of x_i, x_j for all i, j, and no other information about the x_i.
- [x] We use Lagrange multipliers in an optimization problem with inequality ($\leq$) constraints.
- [ ] None of the above

(j) [3 pts] Which of the following distance metrics can be computed exclusively with inner products, assuming $\Phi(x)$ and $\Phi(y)$ are feature mappings of x and y, respectively?

- [ ] $\Phi(x) - \Phi(y)$
- [x] $\|\Phi(x) - \Phi(y)\|_2^2$.
- [ ] $\|\Phi(x) - \Phi(y)\|_1$
- [ ] None of the above

(k) [3 pts] Strong duality holds for:

- [x] Hard Margin SVM
- [ ] Constrained optimization problems in general
- [x] Soft Margin SVM
- [ ] None of the above

(l) [3 pts] Which the following facts about the 'C' in SVMs is (are) true?

- [ ] As C approaches 0, the soft margin SVM is equal to the hard margin SVM
- [ ] A larger C tends to create a larger margin
- [x] None of the above
- [ ] C can be negative, as long as each of the slack variables are nonnegative

Q3. [10 pts] Parameter Estimation

In this problem, we have n trials with k possible types of outcomes $\{1, 2, ..., k\}$. Suppose we observe $X_1, ..., X_k$ where each X_i is the number of outcomes of type i. If p_i refers to the probability that a trial has outcome i, then $(X_1, ..., X_k)$ is said to have a multinomial distribution with parameters $p_1, ..., p_k$, denoted $(X_1, ..., X_k) \sim \text{Multinomial}(p_1, ..., p_k)$. It may be useful to know that the probability mass function of the multinomial distribution is given as follows.

$$P(X_1 = x_1, ..., X_k = x_k) = \frac{n!}{x_1!x_2! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

We want to find the maximum likelihood estimators for $p_1, ..., p_k$. You may assume that $p_i > 0$ for all i.

(a) [4 pts] What is the log-likelihood function, $l(p_1, ..., p_k | X_1, ..., X_k)$?

The likelihood function is given as follows:

$$L(p_1, ..., p_k | X_1, ..., X_k) = P(X_1, ..., X_k | p_1, ..., p_k) = \frac{n!}{X_1!X_2! \dots X_k!} p_1^{X_1} \dots p_k^{X_k}$$

Therefore, the log-likelihood is given as follows:

$$l(p_1, ..., p_k | X_1, ..., X_k) = \log(n!) - \sum_{i=1}^{k} \log(X_i!) + \sum_{i=1}^{n} X_i \log p_i$$

(b) [6 pts] You might notice that unconstrained maximization of this function leads to an answer in which we set each $p_i = \infty$. But this is wrong. We must add a constraint such that the probabilities sum up to 1. Now, we have the following optimization problem.

$$\max_{p_1, ..., p_k} l(p_1, ..., p_k | X_1, ..., X_k)$$
$$\text{s.t.} \sum_{i=1}^{k} p_i = 1$$

Recall that we can use the method of Lagrange multipliers to solve an optimization problem with equality constraints. Using this method, find the maximum likelihood estimators for $p_1, ..., p_k$.

By applying the Method of Lagrange Multipliers, we get the following Lagrangian.

$$L(p_1, ..., p_k, \lambda) = \log(n!) - \sum_{i=1}^{k} \log(X_i!) + \sum_{i=1}^{n} X_i \log p_i + \lambda(1 - \sum_{i=1}^{k} p_i)$$

We take the derivative with respect to each p_i and λ and set it equal to 0.

$$\frac{\partial L}{\partial p_i} = \frac{X_i}{p_i} - \lambda = 0$$
$$\frac{\partial L}{\partial \lambda} = 1 - \sum_{i=1}^{k} p_i = 0$$

Solving for p_i, we get the MLE.

$$\hat{p_i}_{\text{MLE}} = \frac{X_i}{n}$$

Q4. [8 pts] Dual Solution for Ridge Regression

Recall that ridge regression minimizes the objective function:

$$L(w) = \|Xw - y\|_2^2 + \lambda\|w\|_2^2$$

where X is an n-by-d design matrix, w is a d-dimensional vector and y is a n-dimensional vector. We already know that the function $L(w)$ is minimized by

$$w^* = (X^TX + \lambda I)^{-1}X^Ty.$$

Alternatively, the minimizer can be represented by a linear combination of the design matrix rows. That is, there exists a n-dimensional vector α^* such that the objective function $L(w)$ is minimized by $w^* = X^T\alpha^*$. The vector α^* is called the dual solution to the linear regression problem.

(a) [2 pts] Using the relation $w = X^T\alpha$, define the objective function L in terms of α.

$$L(X^T\alpha) = \|XX^T\alpha - y\|_2^2 + \lambda\|X^T\alpha\|_2^2$$

(b) [3 pts] Show that $\alpha^* = (XX^T + \lambda I)^{-1}y$ is a dual solution.

Proof: Using the relation $w = X^T\alpha$, we define the objective function in terms of α

$$\ell(\alpha) = L(X^T\alpha) = \|XX^T\alpha - y\|_2^2 + \lambda\|X^T\alpha\|_2^2.$$

Since $X\alpha^*$ is the minimizer of $L(w)$, the vector α^* should be the minimizer of function $\ell(\alpha)$. By calculating the gradient of function $\ell(\alpha)$ and make it equal to zero, we get

$$XX^T(XX^T\alpha - y + \lambda\alpha) = 0.$$

Since $\ell(\alpha)$ is a convex function, every solution to the above equation is a minimizer of the objective function. Since $\alpha^* = (XX^T + \lambda I)^{-1}y$ is one solution, it establishes the claim.

Alternate Proof: to verify that α^* is a dual solution, it suffices to verify that $w^* = X^T\alpha^*$, or equivalently

$$\begin{aligned}
&(X^TX + \lambda I)^{-1}X^Ty = X^T(XX^T + \lambda I)^{-1}y \\
\Leftrightarrow &X^Ty = (X^TX + \lambda I)X^T(XX^T + \lambda I)^{-1}y \\
\Leftrightarrow &X^Ty = (X^TXX^T + \lambda X^T)(XX^T + \lambda I)^{-1}y \\
\Leftrightarrow &X^Ty = X^T(XX^T + \lambda I)(XX^T + \lambda I)^{-1}y \\
\Leftrightarrow &X^Ty = X^Ty
\end{aligned}$$

which completes the proof.

(c) [3 pts] To make the solution in question (b) well-defined, the matrix $XX^T + \lambda I$ has to be an invertible matrix. Assuming $\lambda > 0$, show that $XX^T + \lambda I$ is an invertible matrix. (Hint: positive definite matrices are invertible)

Since all positive definite matrices are invertible, it suffices to show that the matrix $XX^T + \lambda I$ is positive definite. For any non-zero vector v, we have

$$v^T(XX^T + \lambda I)v = v^TXX^Tv + \lambda v^Tv = (X^Tv)^T(X^Tv) + \lambda v^Tv = \|X^Tv\|_2^2 + \lambda\|v\|_2^2$$

Since $v \neq 0$, we have $\|X^Tv\|_2^2 \geq 0$ and $\lambda\|v\|_2^2 > 0$, which implies that $v^T(XX^T + \lambda I)v > 0$. It establishes that the matrix is positive definite.

Q5. [8 pts] Regularization and Priors for Linear Regression

Linear regression is a model of the form $P(y|\mathbf{x}) \sim \mathcal{N}(\mathbf{w^T x}, \sigma^2)$, where $\mathbf{w}$ is a d-dimensional vector. Recall that in ridge regression, we add an ℓ_2 regularization term to our least squares objective function to prevent overfitting, so that our loss function becomes:

$$J(\mathbf{w}) = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \lambda \mathbf{w^T w} \qquad (*)$$

We can arrive at the same objective function in a Bayesian setting, if we consider a MAP (maximum a posteriori probability) estimate, where $\mathbf{w}$ has the prior distribution $\mathcal{N}(0, f(\lambda, \sigma)I)$.

(a) [3 pts] What is the conditional density of w given the data?

$$P(w|\mathbf{X_i}, Y_i) \propto (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot P(\mathbf{w}) = (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot \prod_{j=1}^{d} P(w_j)$$

(b) [5 pts] What $f(\lambda, \sigma)$ makes this MAP estimate the same as the solution to (*)?

Taking logs, we want to maximize

$$l(\mathbf{w}) = \sum_{i=1}^{n} \log \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2) + \sum_{j=1}^{d} \log P(w_j)$$

$$= \sum_{i=1}^{n} \log(\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2})) + \sum_{j=1}^{d} \log(\frac{1}{\sqrt{2\pi f(\lambda, \sigma)}} exp(\frac{-w_j^2}{2f(\lambda, \sigma)}))$$

$$= -\sum_{i=1}^{n} \frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2} + \frac{-\sum_{j=1}^{d} w_j^2}{2f(\lambda, \sigma)} + n \log(\frac{1}{\sqrt{2\pi}\sigma}) + d \log(\frac{1}{2f(\lambda, \sigma)}),$$

so it is equivalent to minimizing the following function

$$J(\mathbf{w}) = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \frac{\sigma^2}{f(\lambda, \sigma)} \sum_{j=1}^{d} w_j^2$$

$$= \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \lambda \mathbf{w^T w},$$

hence $f(\lambda, \sigma) = \frac{\sigma^2}{\lambda}$.