Midterm Exam
EECS 545: Machine Learning
Winter, 2016

Name:

UM uniqname:

- **Closed book. Three sheets of paper of notes are allowed. No computers, cell phones or calculators.**
  - Showing your work makes partial credit possible.
    If you write nothing at all, it's hard to justify any score but zero.
  - Feel free to use the back of each sheet for scratch paper, but not for credit.
    IMPORTANT: Your exam will be scanned and we will not include the back pages, so do not write anything you want for credit there.
  - Write clearly. If we can't read your writing, it will be marked wrong.
- This course operates under the rules of the College of Engineering Honor Code. Your signature endorses the pledge below. **After** you finish your exam, please sign below:
  I have neither given nor received aid on this examination, nor have I concealed any violations of the Honor Code.

DO NOT WRITE BELOW THIS LINE

Problem	1	2	3	4	5	6	Total
Points							
Max Points	12	6	5	6	8	3	40

Problem 1 (True/False). Are the following statements true or false? Please circle one. (No need for explanations unless you feel the question is ambiguous and want to justify your answer).

(a) The Perceptron algorithm can be interpreted as a type of stochastic gradient descent, assuming the order of the examples is fixed in advance.

(True) False

Indeed, we can write the error function $E(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \max(0, -y_i \mathbf{w}^\top \mathbf{x_i})$. The gradient of the error function with respect to *one* term (i) in the sum is

$$\nabla E_i(\mathbf{w}) = \begin{cases} 0 & \text{when } y_i \mathbf{w}^\top \mathbf{x_i} > 0 \\ -y_i \mathbf{x}_i & \text{otherwise .} \end{cases}$$

This suggests that the Perceptron algorithm can be viewed as a stochastic gradient descent update on the example $\mathbf{x}_i, y_i$.

(b) If you do T sequential updates with the perceptron, you will arrive at the same final $\mathbf{w}$ regardless of the order of the updates.

True (False)

Imagine we are in one dimension, and the examples are $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (1, -1)$, $(x_3, y_3) = (1, 1)$. It is easy to see that we will arrive at the final $w = 1$. But if we reorder the examples so that $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (1, 1)$, $(x_3, y_3) = (1, -1)$, it is easy to check that the final w will be 0.

(c) Linear Discriminant Analysis can only be applied when the dataset in question is linearly separable.

True (False)

There is nothing in the LDA model that requires separability.

(d) The core assumption of Naive Bayes classifiers is that all observed variables (features) are statistically independent.

True (False)

The core assumption of Naive Bayes is that the the observed variables are statistically independent *given the class label!* They are not (necessarily) independent without this conditioning.

(e) Regularized linear regression can be interpreted as the MAP estimate of a model in which the weights w are endowed with a Gaussian prior.

(True) False

Yep. Look at the section entitled "Regularized Least Squares" in Lecture 5.

(f) Naive Bayes is not a generative model.

True (False)

Naive Bayes is based on a model that specifies $\Pr(\mathbf{x}|y, \text{params})$. Hence we can sample (generate) new examples $\mathbf{x}$ from this distribution.

(g) Quadratic discriminant analysis as an approach to classification cannot be **applied** if the true class-conditional density for each class is *not* Gaussian.

`True` (`False`)

The GDA model does make the assumption that the class-conditional densities are multivariate Gaussian distributions. However, the method can be applied for any dataset as long as we can calculate empirical mean and (class-conditional) covariance matrices.

(h) Logistic Regression is a method for solving classification problems.

(`True`) `False`

Don't let the name fool you: Logistic Regression is a method of learning a hypothesis to predict a probability that discriminates between two (or more) classes.

(i) In the least squares regression problem $\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|_2^2$, there may be more than one $\mathbf{w}$ that minimizes this objective. As a result, there may be more than one correct prediction $\hat{\mathbf{y}} = X\mathbf{w}$.

`True` (`False`)

If there is more than one $\mathbf{w}$ minimizing the objective, then the prediction vector $X\mathbf{w}$ will give the same result for each $\mathbf{w}$.

(j) Suppose that we have a convex function $f(\mathbf{x})$ defined for $\mathbf{x} \in \mathbb{R}^n$. Then $g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$ is also convex for any $A \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$.

(`True`) `False`

Convexity is always preserved under affine transformations of the input.

(k) The optimization problem for hard-margin SVM always has at least one feasible solution for any training dataset.

True (False)

The hard-margin SVM has no feasible solution when the data are not linearly separable.

(l) Let $k(\cdot, \cdot)$ be a valid (positive semi-definite) kernel mapping $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Let $M \in \mathbb{R}^{d \times d}$ be any matrix, and define the new function $\tilde{k}(\mathbf{u}, \mathbf{v}) := k(M\mathbf{u}, M\mathbf{v})$. Then $\tilde{k}$ is also a valid PSD kernel.

(True) False

The function $\tilde{k}$ is clearly symmetric. To check that it is a PSD kernel, let us imagine a sample of examples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and consider the matrix

$$M = [\tilde{k}(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1\ldots n} = [k(M\mathbf{x}_i, M\mathbf{x}_j)]_{i,j=1\ldots n}.$$

We know that k is a PSD kernel, so the matrix $M = [k(M\mathbf{x}_i, M\mathbf{x}_j)]_{i,j=1\ldots n}$ must be PSD by definition. We can thus conclude that $\tilde{k}$ is a PSD kernel as well.

Problem 2 (MLE for the Uniform Distribution). Consider a uniform distribution centered on 0 with width $2a$. The density function is given by

$$P(x) = \begin{cases} \frac{1}{2a} & \text{if } x \in [-a, a] \\ 0 & \text{otherwise} \end{cases}$$

(a) Given a dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, what is the maximum likelihood estimate $\hat{a}_{ML}$ of the quantity a?

(1) Given data $\mathcal{X} = \{x_1, \ldots, x_n\}$, the likelihood function is zero unless all training points $\mathcal{X}$ lie in the range $[-a, a]$. Mathematically,

$$\ell(a|\mathcal{X}) = P(\mathcal{X}|a) = \prod_{k=1}^{n} P(x_k|a) = \begin{cases} \left(\frac{1}{2a}\right)^n & \text{if } \mathcal{X} \subset [-a, a] \\ 0 & \text{otherwise} \end{cases}$$

(2) Clearly, we need $\hat{a}_{ML} \geq \max_k |x_k|$. The likelihood function *decreases* as a increases, so the likelihood is maximum when we choose a to be as small as possible. Therefore,

$$\boxed{\hat{a}_{ML} = \max_k |x_k|}$$

.

(b) What probability would the model assign to a new data point x_{n+1} using $\hat{a}_{ML}$?

$$P(x_{n+1}|\hat{a}_{ML}) = \begin{cases} \frac{1}{2\hat{a}_{ML}} & \text{if } x \in [-\hat{a}_{ML}, \hat{a}_{ML}] \\ 0 & \text{otherwise} \end{cases}$$

Problem 3 (Logistic Regression). We described the probability model behind logistic regression as follows. First we assume a parameter vector $\mathbf{w} \in \mathbb{R}^d$ is given, and for an arbitrary input $\mathbf{x} \in \mathbb{R}^d$ the target y is chosen as

$$y \sim \text{Bernoulli}(\sigma(\mathbf{w}^\top \mathbf{x})).$$

Recall that $\sigma(\cdot)$ is the standard sigmoid function and $\text{Bernoulli}(\sigma(\mathbf{w}^\top \mathbf{x}))$ is the distribution that returns 1 with probability $\sigma(\mathbf{w}^\top \mathbf{x})$ and 0 with probability $1 - \sigma(\mathbf{w}^\top \mathbf{x})$.

We can express the model for Logistic Regression in another way:

$$\epsilon \sim \text{FancyDistribution}$$
$$y = \mathbb{I}[\epsilon + \mathbf{w}^\top \mathbf{x} \geq 0]$$

where $\mathbb{I}[\cdot]$ returns 1 when the input is True, and 0 when the input is False. What is the distribution FancyDistribution? Please describe this distribution in terms of its probability density function (PDF). (Partial credit will be given for only providing the cumulative distribution function (CDF).)

We need to choose a distribution over ϵ so that

$$\Pr(\epsilon + \mathbf{w}^\top \mathbf{x} \geq 0) = \sigma(\mathbf{w}^\top \mathbf{x})$$
$$\Pr(\epsilon + \mathbf{w}^\top \mathbf{x} < 0) = 1 - \sigma(\mathbf{w}^\top \mathbf{x}).$$

If we let $y = -\mathbf{w}^\top \mathbf{x}$, and we use the second equation, we see that

$$\Pr(\epsilon + \mathbf{w}^\top \mathbf{x} < 0) = \Pr(\epsilon < y) = 1 - \sigma(-y) = \sigma(y),$$

where the last equality can be checked easily.

In other words, we have just stated that the CDF of the random variable ϵ should be *exactly* equal to the sigmoid function. Recall that for a smooth distribution, the PDF of a random variable is the derivative of the CDF. If we let $f(y)$ be the PDF of ϵ then we have

$$f(y) = \frac{d}{dy}\sigma(y) = \frac{d}{dy}\left(\frac{1}{1+\exp(-y)}\right) = \frac{\exp(-y)}{(1+\exp(-y))^2}$$

Problem 4 (Convexity). Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Consider the function

$$f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T A \mathbf{x} + 2\mathbf{x}^T B \mathbf{y} + \mathbf{y}^T C \mathbf{y}$$

for symmetric matrices $A \in \mathbb{R}^{n \times n}$ and $C \in \mathbb{R}^{m \times m}$, and some matrix $B \in \mathbb{R}^{n \times m}$. Define $\mathbf{z}$ as the concatenation of $\mathbf{x}$ and $\mathbf{y}$; i.e., $\mathbf{z} = [x_1, \ldots, x_n, y_1, \ldots, y_m]^T$.

(a) What is the Hessian $\nabla^2 f(\mathbf{z})$? (That is, the Hessian with respect to the vector $\mathbf{z}$)

If we consider $\mathbf{y}$ fixed, we know already that

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = 2A$$

and, similarly, if $\mathbf{x}$ is fixed,

$$\nabla^2_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 2C$$

Now we just need $\nabla_{\mathbf{x}} \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$:

$$\begin{aligned} \nabla_{\mathbf{x}} \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) &= \nabla_{\mathbf{x}} \left(2B^T \mathbf{x} + 2C\mathbf{y} \right) \\ &= 2B \end{aligned}$$

And, naturally, $\nabla_{\mathbf{y}} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = (\nabla_{\mathbf{x}} \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))^T$. So

$$\begin{aligned} \nabla^2_{\mathbf{z}} f(\mathbf{z}) &= \begin{bmatrix} \nabla^2_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) & \nabla_{\mathbf{x}} \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \\ \nabla_{\mathbf{y}} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) & \nabla^2_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{bmatrix} \\ &= 2 \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \end{aligned}$$

(b) Assume A and C are positive semi-definite. Give a single choice for the matrix B which guarantees that $f(\mathbf{z})$ is convex in $\mathbf{z}$. (Note that several choices for B are possible.)

(For extra credit, can you completely characterize all matrices B for which $f(\mathbf{z})$ is convex? Feel free to define new matrices as necessary. *Hint: Every PSD matrix $M \in \mathbb{R}^{d\times d}$ can be written as a Gram matrix, i.e. $M = X^T X$ for some (non-unique) matrix $X \in \mathbb{R}^{p\times d}$, for any $p \geq d$.*)

Since $f(\mathbf{z})$ is twice differentiable, its Hessian being positive semi-definite is equivalent to f being convex. The Hessian must be able to be written as a Gram matrix, then, so we can write it as

$$\begin{aligned}\nabla^2_{\mathbf{z}} f(\mathbf{z}) &= 2 \begin{bmatrix} U^T \\ V^T \end{bmatrix} \begin{bmatrix} U & V \end{bmatrix} \\ &= 2 \begin{bmatrix} U^T U & U^T V \\ V^T U & V^T V \end{bmatrix}\end{aligned}$$

Where $U \in \mathbb{R}^{p\times n}$, $V \in \mathbb{R}^{p\times m}$, and $p \geq m + n$. Then we must have that $A = U^T U$, $B = U^T V$, and $C = V^T V$. This is the general condition, then: that $B = U^T V$ for some matrices U, V such that $A = U^T U$ and $C = V^T V$.

A special, and obvious with respect to making f convex, case of this is $B = 0$. Observe that if we multiply U or V on the left by an orthogonal matrix, the result is an equally valid choice of U or V, respectively. Then, if we choose a sufficiently large p (e.g., $p \geq 2\max(n, m)$), then we can rotate U and V until they span non-overlapping subspaces (in other words, the columns of U are orthogonal to the columns of V), we get $B = 0$.

Problem 5 (Lagrangian). The dual norm of a norm $\|\cdot\|$ is denoted $\|\cdot\|_*$ and is defined as

$$\|\mathbf{x}\|_* \triangleq \max_{\mathbf{z}:\|\mathbf{z}\|\leq 1} \mathbf{x}^T\mathbf{z}$$

In this problem, let the norm $\|\cdot\|$ be the ℓ^p norm; i.e., for $p \geq 1$,

$$\|\mathbf{x}\| = \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

(a) Write down the newpagen for the dual norm optimization problem. This should be a function that you need to minimize. *Hint:* $\|\mathbf{x}\|_p \leq 1 \iff \|\mathbf{x}\|_p^p \leq 1$.

We need to minimize $-\mathbf{x}^T\mathbf{z}$, and we have the constraint that $\sum_{i=1}^{n} |z_i|^p - 1 \leq 0$, so the Lagrangian is

$$\mathcal{L}(\mathbf{z}, \lambda) = -\mathbf{x}^T\mathbf{z} + \lambda\left(\sum_{i=1}^{n} |z_i|^p - 1\right)$$

(b) Obtain the Lagrange dual function $\mathcal{L}_D(\cdot)$. The primal variable should not appear in this expression. You do not need to solve the Lagrange dual problem.

The Lagrangian is convex in $\mathbf{z}$, so we can find the minimum with differentiation:

$$\frac{\partial \mathcal{L}(\mathbf{z}, \lambda)}{\partial z_i} = -x_i + \lambda p |z_i|^{p-1} \text{sign}(z_i) = 0$$

$$x_i \text{sign}(z_i) = \lambda p |z_i|^{p-1}$$

Since the right hand side is always non-negative, we must have that $\text{sign}(z_i) = \text{sign}(x_i)$. So at the minimum, we have that for the optimal $\mathbf{z}^*$,

$$|z_i^*| = \left(\frac{|x_i|}{\lambda p} \right)^{\frac{1}{p-1}}$$

The dual function is

$$\begin{aligned}
\mathcal{L}_D(\lambda) &= -\mathbf{x}^T \mathbf{z}^* + \lambda \left(\sum_{i=1}^n |z_i^*|^p - 1 \right) \\
&= \sum_{i=1}^n \left(-|x_i||z_i^*| + \lambda |z_i^*|^p \right) - \lambda \\
&= \sum_{i=1}^n \left(-p\lambda |z_i^*|^p + \lambda |z_i^*|^p \right) - \lambda \\
&= \lambda(1-p) \sum_{i=1}^n \lambda |z_i^*|^p - \lambda \\
&= \lambda(1-p) \sum_{i=1}^n \left(\frac{|x_i|}{\lambda p} \right)^{\frac{p}{p-1}} - \lambda \\
&= \lambda^{\frac{1}{1-p}} (1-p) \sum_{i=1}^n \left(\frac{|x_i|}{p} \right)^{\frac{p}{p-1}} - \lambda
\end{aligned}$$

We didn't ask for you to solve the dual problem, but to do so, differentiate with respect to λ:

$$\frac{\partial \mathcal{L}_D(\lambda)}{\partial \lambda} = \lambda^{\frac{p}{1-p}} \sum_{i=1}^n \left(\frac{|x_i|}{p} \right)^{\frac{p}{p-1}} - 1 = 0$$

Letting $q = p/(p-1)$,

$$\lambda^* = \frac{1}{p} \left(\sum_{i=1}^n |x_i|^q \right)^{\frac{1}{q}} = \frac{1}{p} \|\mathbf{x}\|_q$$

This gives the value of the dual function at the maximum:

$$\begin{aligned}
d^* &= \lambda^* \left((\lambda^* p)^{-q} (1-p) \|\mathbf{x}\|_q^q - 1 \right) \\
&= \frac{1}{p} \|\mathbf{x}\|_q (-p) \\
&= -\|\mathbf{x}\|_q
\end{aligned}$$

Strong duality holds here (there is only one constraint and the original problem was convex, so it is sufficient), so the dual norm of the ℓ^p norm is

$$\|\mathbf{x}\|_{p*} = -d^* = \|\mathbf{x}\|_q$$

where $1/p + 1/q = 1$.

Problem 6 (Challenge Problem). Suppose we are in a situation where we observe a function $f(\cdot)$, and we want to approximate this function on the interval $[a, b]$ using a linear combination of K linearly independent functions $g_1(\cdot), \ldots, g_K(\cdot)$. We will use least squares error as our objective function; i.e., we want to solve

$$\min_{\mathbf{w} \in \mathbb{R}^K} \int_a^b \left(f(x) - \sum_{k=1}^K w_k g_k(x) \right)^2 dx$$

Give an expression for the $\mathbf{w}$ that solves this problem. Feel free to define new vectors and matrices as necessary. *Hint: Define the inner product for two functions f and g as $\langle f, g \rangle \triangleq \int_a^b f(x)g(x)dx$, and assume we can compute this value between all pairs from $f, g_1, g_2, \ldots, g_K$.*

Let's expand the squared term:

$$\min_{\mathbf{w} \in \mathbb{R}^K} \int_a^b f(x)^2 - 2f(x) \sum_{k=1}^K w_k g_k(x) + \sum_{k,\ell=1}^K w_k w_\ell g_k(x) g_\ell(x) dx$$

By linearity of integration, we have

$$\min_{\mathbf{w} \in \mathbb{R}^K} \langle f, f \rangle - 2 \sum_{k=1}^K w_k \langle f, g_k \rangle + \sum_{k,\ell=1}^K w_k w_\ell \langle g_k, g_\ell \rangle$$

If we define the matrix $G = [\langle g_k, g_\ell \rangle]_{k,\ell=1}^K$ and the vector $F = [\langle f, g_k \rangle]_{k=1}^K$, our problem is simply

$$\min_{\mathbf{w} \in \mathbb{R}^K} \langle f, f \rangle - 2F^T \mathbf{w} + \mathbf{w}^T G \mathbf{w}$$

Differentiation with respect to $\mathbf{w}$ gives

$$-2F + 2G\mathbf{w} = 0$$

So any solution to the following equation is valid

$$G\mathbf{w} = F$$

If G is invertible (which is true if g_k are linearly independent with respect to $[a, b]$), the unique solution is

$$\mathbf{w} = G^{-1} F$$