DS-GA 1003 Machine Learning and Computational Statistics, Spring 2015
Midterm Exam

Prof. David S. Rosenberg

April 8, 2015

Instructions: **Please write your NYU NetID at the top of each page of your exam.** Write your solutions in the space provided below each question. If you need additional space, you may use the extra blank sheet at the end of the test. If you need even more pages, you may ask the instructor for additional paper. You may also write in any white space on the test, just be very clear about what should be graded and what should be ignored. Note that there is an appendix to the test giving some results on convex functions.

Name:			
NYU NetID:			
Q	**Name**	**Max Score**	**Score**
1	**True/False Questions**	**10**	
2	**Short Answer**	**17**	
3	**Hypothesis Spaces**	**6**	
4	**Kernelizing Ridge Regression**	**13**	
5	**Ivanov and Tikhonov Regularization**	**14**	
6	**Square Hinge Loss**	**15**	
7	**Conditional Exponential Distributions**	**15**	
	Total	**90**	

1 True / False Questions

1. (**True or False**, 1 pt) When using (unregularized) linear regression, adding new features always improves the performance on training data, or at least never make it worse.
2. (**True or False**, 1 pt)When using a (unregularized) linear regression, adding new features always improves the performance on test data, or at least never make it worse.
3. (**True or False**, 1 pt) Overfitting is more likely when the set of training data is small.
4. (**True or False**, 1 pt) Overfitting is more likely when the hypothesis space is small.
5. (**True or False**, 1 pt) Approximation error decreases to zero as the amount of training data goes to infinity.
6. (**True or False**, 1 pt) If the empirical risk function is not convex, more training data may not help estimation error.
7. (**True or False**, 1 pt) If a decision tree is trained on data for which two features are exactly equal, the resulting tree will be the same whether or not we remove one of those two features.
8. (**True or False**, 1 pt) Suppose we fit Lasso regression to a data set. If we rescale one of the features by multiplying it by 10, and we then refit Lasso regression with the same regularization parameter, then it is more likely for that feature to be excluded from the model
9. (**True or False**, 1 pt) Adaboost with decision stumps will eventually reach zero training error, provided we run enough rounds of boosting.
10. (**True or False**, 1 pt) When you have a very large data set of size n, which is much larger than the dimension d of the feature space, kernel methods are probably not a good idea.

2 Short Answer

1. (1 pt) Circle all of the loss functions that may lead to support vectors: **exponential loss, hinge loss, squared hinge loss, logistic loss, square loss**.
2. (4 pts) We have a dataset $\mathcal{D} = \{(0,1), (1,4), (2,3)\}$ that we fit by minimizing an objective function of the form:

$$J(\alpha_0, \alpha_1) = \frac{1}{3}\sum_{i=1}^{3} (\alpha_0 + \alpha_1 x_i - y_i)^2 + \lambda_1 (\alpha_0 + \alpha_1) + \lambda_2(\alpha_0^2 + \alpha_1^2),$$

   and the corresponding fitted function is given by $f(x) = \alpha_0 + \alpha_1 x$. We tried four different settings of λ_1 and λ_2, and the results are shown in Figure 2.1. For each of the following parameter settings, give the number of the plot that shows the resulting fit.

   (a) (1 pt) $\lambda_1 = 0$ and $\lambda_2 = 0$.
   (b) (1 pt) $\lambda_1 = 5$ and $\lambda_2 = 0$.
   (c) (1 pt) $\lambda_1 = 0$ and $\lambda_2 = 10$.
   (d) (1 pt) $\lambda_1 = 0$ and $\lambda_2 = 2$.

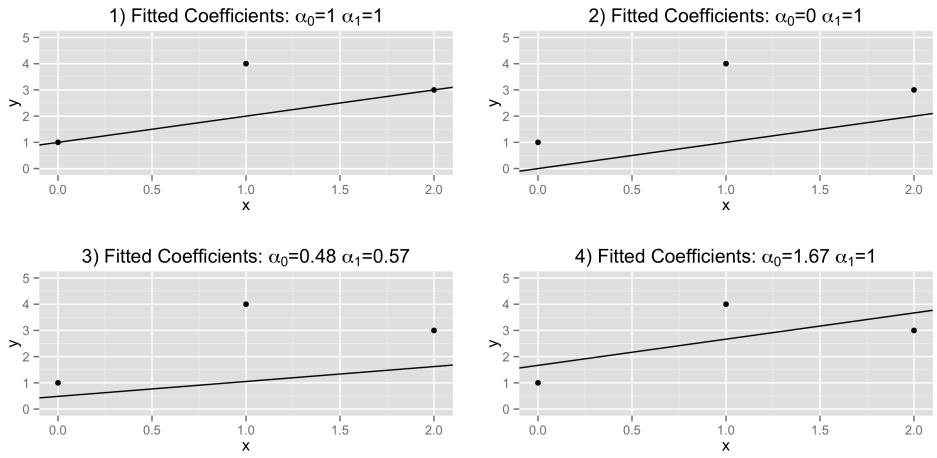

Figure 2.1: Linear fits with different penalizations.

3. (2 pts) Show that the following kernel function is a Mercer kernel (i.e. it represents an inner product):

$$k(x, y) = \frac{x^T y}{\|x\|\|y\|},$$

where $x, y \in \mathbf{R}^d$.

4. (2 pts) Consider the binary classification problem shown in Figure 2.2: Denote the input space by

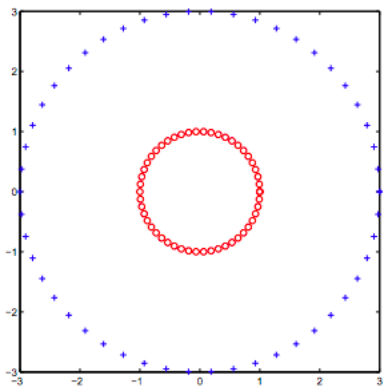

Figure 2.2: For a short-answer problem.

$\mathcal{X} = \{(x_1, x_2) \in \mathbf{R}^2\}$. Give a feature mapping for which a linear classifier could perfectly separate the two classes shown.

5. (2 pts) For the classification problem in Figure (2.2), circle all classifiers that could perfectly separate the classes: **linear SVM, SVM with quadratic kernel, decision stumps (i.e. classification trees with only two leaf nodes), AdaBoost with decision stumps, SVM with radial basis function kernel**.

6. (2 pts) Let $\mathcal{F}_1 = \{\text{binary classification trees of depth 2}\}$. Let $\mathcal{F}_2 = \{\text{all linear classifiers}\}$. Draw a binary classification dataset for which a member of $\mathcal{F}_1$ can perfectly separate the data, while no member of $\mathcal{F}_2$ can. Show the splits and the decision boundary for the tree.

7. (2 pts) Same $\mathcal{F}_1$ and $\mathcal{F}_2$ as in the previous problem. Draw a binary classification dataset for which a member of $\mathcal{F}_2$ can perfectly separate the data, while no member of $\mathcal{F}_1$ can.

8. (2 pts) Suppose we fit a hard-margin SVM to N data points, and we have 2 data points "on the margin". If we add a new data point to the training set and refit the SVM, what's the largest number of data points that could end up "on the margin". Support your answer (a picture could suffice).

3 Hypothesis Spaces

1. (2 pt) Consider the following two hypothesis spaces:

$$\mathcal{F}_1 = \{f(x) = e^{w_1}x + w_2 x \mid w_1, w_2 \in \mathbf{R}\} \qquad \mathcal{F}_2 = \{f(x) = wx \mid w \in \mathbf{R}\}$$

Suppose we are selecting hypotheses using empirical risk minimization (without any penalty). Are there any situations in which one of these hypothesis spaces would be preferred to the other? Why?

2. (2 pt) Same question, with the following hypothesis spaces:

$$\mathcal{F}_1 = \{f(x) = e^{w_1}x \mid w_1 \in \mathbf{R}\} \qquad \mathcal{F}_2 = \{f(x) = wx \mid w \in \mathbf{R}\}$$

3. (2 pt) Same question, with the following hypothesis spaces:

$$\mathcal{F}_1 = \{\text{trees of depth at most 2}\} \qquad \mathcal{F}_2 = \{\text{trees with at most 4 leaf nodes}\}$$

4 Kernelizing Ridge Regression

Suppose our input space is $\mathcal{X}$=$\mathbf{R}^d$ and our output space is $\mathcal{Y} = \mathbf{R}$. Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set from $\mathcal{X} \times \mathcal{Y}$. We'll use the "design matrix" $X \in \mathbf{R}^{n \times d}$, which has the input vectors as rows:

$$X = \begin{pmatrix} -x_1- \\ \vdots \\ -x_n- \end{pmatrix}.$$

Recall the ridge regression objective function:

$$J(w) = ||Xw - y||^2 + \lambda||w||^2,$$

for $\lambda > 0$.

1. (4 pts) Derive a closed form expression for the minimizer of $J(w)$.

2. (2 pts) Show that w^*, the minimizer of $J(w)$, can be written as $w = X^T\alpha$, where $\alpha = \lambda^{-1}(y - Xw)$. [If you don't remember the trick, you may want to leave this problem until the end. The rest of the problems do not depend on this one.]

3. (1 pt) Based on the fact that $w = X^T\alpha$, explain why we say w is "in the span of the data."

4. (3 pts) Give a kernelized expression for α in terms of the kernel matrix XX^T. (Hint: Plug in $w = X^T\alpha$ to the expression for α.)

5. (2 pts) Give a kernelized expression for Xw, the predicted values on the training points.

6. (1 pt) Give a kernelized expression for the prediction on new points, stored in the matrix X_P, where

$$X_P = \begin{pmatrix} -x_1- \\ \vdots \\ -x_n- \end{pmatrix}.$$

5 Ivanov and Tikhonov Regularization

In lecture there was a claim that the Ivanov and Tikhonov forms of ridge and lasso regression are equivalent. We will now prove a much more general result.

5.1 Tikhonov optimal implies Ivanov optimal

Let $\phi : \mathcal{F} \to \mathbf{R}$ be any performance measure of $f \in \mathcal{F}$ and let $\Omega : \mathcal{F} \to \mathbf{R}$ be any complexity measure. For example, for ridge regression over the linear hypothesis space $\mathcal{F} = \left\{ f_w(x) = w^T x \mid w \in \mathbf{R}^d \right\}$, we would have $\phi(f_w) = \frac{1}{n} \sum_{i=1}^{n} \left(w^T x_i - y_i \right)^2$ and $\Omega(f_w) = w^T w$.

1. (3 pts) Suppose that for some $\lambda > 0$ we have the Tikhonov regularization solution

$$f^* = \arg\min_{f \in \mathcal{F}} \left[\phi(f) + \lambda \Omega(f) \right]. \tag{5.1}$$

Show that f^* is also an Ivanov solution. That is, $\exists r > 0$ such that

$$f^* = \arg\min_{f \in \mathcal{F}} \phi(f) \text{ subject to } \Omega(f) \leq r. \tag{5.2}$$

(Hint: Start by figuring out what r should be. Then one approach is proof by contradiction: suppose f^* is not the optimum in (5.2) and show that contradicts the fact that f^* solves (5.1).)

5.2 Ivanov optimal implies Tikhonov optimal

For the converse, we will restrict our hypothesis space to a parametric set. That is, $\mathcal{F} = \left\{ f_w(x) : \mathcal{X} \to \mathbf{R} \mid w \in \mathbf{R}^d \right\}$. So we will now write ϕ and Ω as functions of $w \in \mathbf{R}^d$.

Let w^* be a solution to the following Ivanov optimization problem:

$$\begin{array}{ll} \text{minimize} & \phi(w) \\ \text{subject to} & \Omega(w) \leq r. \end{array}$$

Assume that strong duality holds for this optimization problem and that the dual solution is attained. Then we will show that there exists a $\lambda \geq 0$ such that $w^* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda \Omega(w) \right]$.

1. (1 pt) Write the Lagrangian $L(w, \lambda)$ for the Ivanov optimization problem.

2. (2 pts) Write the dual optimization problem in terms of the dual objective function $g(\lambda)$, and give an expression for $g(\lambda)$. [Writing $g(\lambda)$ as an optimization problem is expected - don't try to solve it.]

3. (4 pts) We assumed that the dual solution is attained, so let $\lambda^* = \arg\max_{\lambda \geq 0} g(\lambda)$. We also assumed strong duality, which implies $\phi(w^*) = g(\lambda^*)$. Show that the minimum in the expression for $g(\lambda^*)$ is attained at w^*. [Hint: You can use the same approach we used when we derived that strong duality implies complementary slackness.] **Conclude the proof** by showing that for the choice of $\lambda = \lambda^*$, we have $w^* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda\Omega(w)\right]$.

5.3 Ivanov implies Tikhonov for Ridge Regression.

To show that Ivanov implies Tikhonov for the ridge regression problem (square loss with ℓ_2 regularization), we need to demonstrate strong duality and that the dual optimum is attained. Both of these things are implied by Slater's constraint qualifications.

1. (4 pts) Show that the Ivanov form of ridge regression is a convex optimization problem with a strictly feasible point.

6 Square Hinge Loss and Huberized Square Hinge Loss

The squared hinge loss is a margin loss given by

$$\ell(m) = \left[(1-m)_+\right]^2,$$

where $(m)_+ = m1(m > 0)$ is the "positive part" of m.

1. (2 pts) Suppose we have a training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathcal{X} = \mathbf{R}^d$ and $y_i \in \mathcal{Y} = \{-1, 1\}$, for all $i = 1, \ldots, n$. Consider the linear hypothesis space $\mathcal{F} = \left\{f(x) = w^T x \mid w \in \mathbf{R}^d\right\}$. Write the objective function $J(w)$ for ℓ_2-regularized empirical risk minimization with the square hinge loss over the space $\mathcal{F}$, where $\mathcal{F}$ is parameterized by w.

2. (2 pts) It turns out that $J(w)$ is differentiable at every w. Give the derivative of $J(w)$.

3. (3 pts) Give pseudocode or otherwise explain how you would use stochastic gradient descent to find $w^* = \arg\min_w J(w)$. You need to specify your approach to the step size, but you do not have to specify a stopping criterion, though you may if you like.

4. (2 pts) Justify the claim that the output of SGD can be written in the form:

$$w = \sum_{i=1}^{n} \beta_i x_i.$$

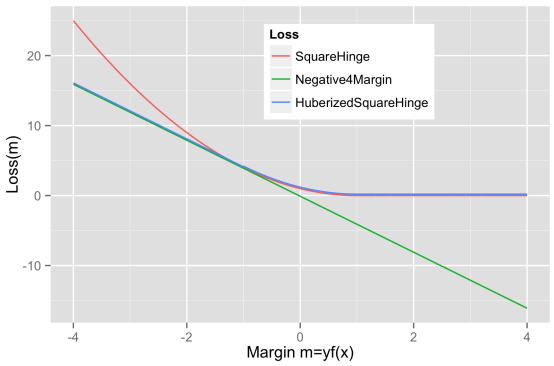

Figure 6.1: Some margin losses.

5. (2 pts) In relation to the SGD algorithm, how would you characterize the x_i's that are support vectors?

6. (2 pts) Show that $J(w)$ is convex. (You are free to cite any of the facts given in Appendix A.)

7. (2 pts) The "Huberized" square hinge loss (shown in Figure 6.1) is a margin loss given by

$$\ell(m) = \begin{cases} -4m & m < -1 \\ \left[(1-m)_+\right]^2 & \text{otherwise.} \end{cases}$$

When might you prefer the Huberized square hinge loss to the square hinge loss?

7 Conditional Exponential Distributions

Suppose we want to model the amount of time one will have to wait for a taxi pickup based on the location and the time. The exponential distribution is a natural candidate for this situation. The exponential distribution is a continuous distribution supported on $[0, \infty)$. The set of all exponential probability density functions is given by

$$\text{ExpDists} = \left\{ p_\lambda(y) = \lambda e^{-\lambda y} 1(y \in [0, \infty)) \mid \lambda \in (0, \infty) \right\}.$$

Recall that a family is a **natural exponential family** of continuous distributions on $\mathbf{R}$ with parameter $\theta \in \mathbf{R}$ if its densities can be written as

$$p_\theta(y) = \frac{1}{Z(\theta)} h(y) \exp\left[\theta y\right],$$

where $Z(\theta) = \int h(y) \exp\left[\theta y\right] \, dy$ is the **partition function**. θ is called the **natural parameter**, and the **natural parameter space** Θ consists of all θ for which $Z(\theta) < \infty$. $h(y)$ is called the **base measure.**

1. (4 pts) Write the family of exponential distributions as a natural exponential family. Give expressions for the base measure and the partition function. Identify the natural parameter space.

2. (3 pts) Let $x \in \mathbf{R}^d$ represent the input features from which we want to predict an exponential distribution. We will use a generalized linear model (GLM) approach. Suggest a reasonable function ψ to map $w^T x$ to a value in the natural parameter space Θ. Then write an expression for $p_w(y \mid x)$, the predicted probability density function conditioned on x.

3. (3 pts) Suppose we have a data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbf{R}^d$ and $y_i \in [0, \infty)$ for $i = 1, \ldots, n$. Give the optimization problem you would solve to fit the GLM we have been discussing to training data $\mathcal{D}$.

4. (5 pts) Suppose we think that a linear function of x doesn't extract enough information, and we'd like to use a more expressive model. For full credit, explain how you would use gradient boosting in this situation. For partial credit, present another reasonable approach to this problem.

A Convexity

A.1 Examples of Convex Functions (BV 3.1.5)

Functions mapping from $\mathbf{R}$:

- $x \mapsto e^{ax}$ is convex on $\mathbf{R}$ for all $a \in \mathbf{R}$
- $x \mapsto x^a$ is convex on $\mathbf{R}_{++}$ when $a \geq 1$ or $a \leq 0$ and concave for $0 \leq a \leq 1$
- $|x|^p$ for $p \geq 1$ is convex on $\mathbf{R}$
- $\log x$ is concave on $\mathbf{R}^{++}$
- $x \log x$ (either on $\mathbf{R}_{++}$ or on $\mathbf{R}_+$ if we define $0 \log 0 = 0$) is convex

Functions mapping from $\mathbf{R}^n$:

- Every norm on $\mathbf{R}^n$ is convex
- Max: $(x_1, \ldots, x_n) \mapsto \max\{x_1 \ldots, x_n\}$ is convex on $\mathbf{R}^n$

A.2 Operations the preserve convexity (BV 3.2, p. 79)

A.2.1 Nonnegative weighted sums

If $f_1, \ldots, f_m$ are convex and $w_1, \ldots, w_m \geq 0$, then $f = w_1 f_1 + \cdots + w_m f_m$ is convex. is convex in x (provided the integral exists).

A.2.2 Composition with an affine mapping

A function $f : \mathbf{R}^n \to \mathbf{R}^m$ is an **affine function** (or **affine mapping)** if it is a sum of a linear function and a constaint. That is, if it has the form $f(x) = Ax + b$, where $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$.

Composition of a convex function with an affine function is convex. More precisely: suppose $f : \mathbf{R}^n \to \mathbf{R}$, $A \in \mathbf{R}^{n \times m}$ and $b \in \mathbf{R}^n$. Define $g : \mathbf{R}^m \to \mathbf{R}$ by $g(x) = f(Ax + b)$, with $\mathrm{dom} g = \{x \mid Ax + b \in \mathrm{dom} f\}$. Then if f is convex, then so is g; if f is concave, so is g. If f is **strictly** convex, and A has linearly independent columns, then g is also strictly convex.

A.2.3 Simple Composition Rules

- If g is convex then $\exp g(x)$ is convex.
- If g is convex and nonnegative and $p \geq 1$ then $g(x)^p$ is convex.
- If g is concave and positive then $\log g(x)$ is concave
- If g is concave and positive then $1/g(x)$ is convex.

A.2.4 Maximum of convex functions is convex (BV Section 3.2.3, p. 80)

Note: Below we use this to prove that the Lagrangian dual function is concave.

If $f_1, \ldots, f_m : \mathbf{R}^n \to \mathbf{R}$ are convex, then their pointwise maximum

$$f(x) = \max\{f_1(x), \ldots, f_m(x)\}$$

is also convex with domain $\mathrm{dom} f = \mathrm{dom} f_1 \cap \cdots \cap \mathrm{dom} f_m$.

This result extends to the supremum over arbitrary sets of functions (including uncountably infinite sets).