DS-GA 1003: Machine Learning and Computational Statistics
Spring 2016
One-Hour Exam

Prof. David S. Rosenberg

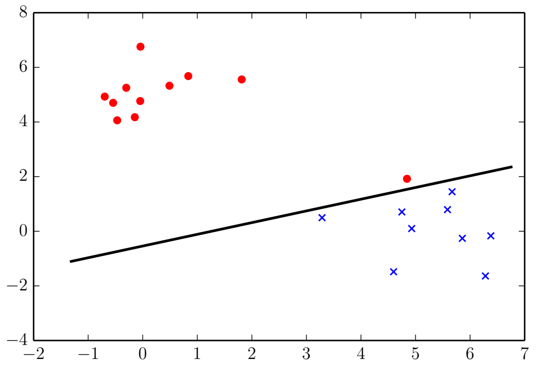
March 3, 2016

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. 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.

Name:			
NYU NetID:			
Q	**Name**	**Max Score**	**Score**
1	**Separable Data**	**3**	
2	**Step Directions in Optimization**	**3**	
3	**Ridge Regression**	**2**	
4	**Perceptron**	**3**	
5	**Regularized Perceptron**	**9**	
6	**Regularization and the Bias Term**	**6**	
	Total	**26**	

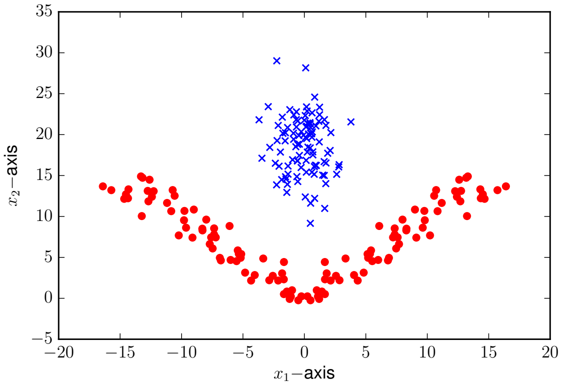
1 Separable Data

1. [1] Draw the maximum margin hyperplane that separates the two classes in the image below. [The two classes are indicated with circles and x's].



SOLUTION: Hyperplane separates the two classes.

1. [2] The picture below shows two classes below in the feature space. Referring to the x-axis as x_1 and the y-axis as x_2, what feature or features could we add that would make this data linearly separable? [Note that the two classes are indicated with circles and x's]



SOLUTION: Reflect the x_1-axis: $(|x_1|, x_2)$

2 Step Directions in Optimization

1. [1] $J(w) : \mathbf{R}^d \to \mathbf{R}$ is a **convex**, **differentiable** objective function with minimizer w^*. Assume w_i is not a minimizer of $J(w)$. Let $g = \nabla J(w_i)$, and let $w_{i+1} = w_i - \eta g$, for some $\eta > 0$. Circle all statements below that we know will be true for small enough $\eta > 0$ (circling none of them is allowed):

   (a) $J(w_{i+1}) < J(w_i)$

   (b) $\|w^* - w_{i+1}\| < \|w^* - w_i\|$
   SOLUTION: (a) is true be definition of gradient and (b) is true because a subgradient step takes us closer to the minimizer (proved in slides).

2. [1] $J(w) : \mathbf{R}^d \to \mathbf{R}$ is a **convex** objective function with minimizer w^*. Assume w_i is not a minimizer of $J(w)$. Let $g \in \partial J(w_i)$ be a **subgradient** of J at w_i, and let $w_{i+1} = w_i - \eta g$, for some $\eta > 0$. Circle all statements below that we know will be true for small enough $\eta > 0$ (circling none of them is allowed):

   (a) $J(w_{i+1}) < J(w_i)$

   (b) $\|w^* - w_{i+1}\| < \|w^* - w_i\|$
   SOLUTION: Just (b), by reason above. Subgradient step doesn't necessarily decrease objective function value (also discussed in slides).

3. [1] Let $J(w) = \frac{1}{n}\sum_{i=1}^{n} J_i(w)$, where each $J_i(w) : \mathbf{R}^d \to \mathbf{R}$ is **convex and differentiable.** Suppose $J(w)$ has minimizer w^*. Let $g = \nabla J_1(w)$. [Please **note** the subscript on J.] Let $w_{i+1} = w_i - \eta g$, for some $\eta > 0$. Circle all statements below that we know will be true for small enough $\eta > 0$ (circling none of them is allowed):

   (a) $J(w_{n+1}) < J(w_n)$

   (b) $\|w^* - w_{n+1}\| < \|w^* - w_n\|$
   SOLUTION: Neither. This is a stochastic gradient step, for which we have no guarantee about the change of any individual step. Over the long term, SGD eventually takes us to the minimizer under some conditions.

3 Ridge Regression

We are given a training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbf{R}^d \times \mathbf{R}$. We put the x's together into a "design matrix" X, where x_i is the ith row of X. So $X \in \mathbf{R}^{n \times d}$. With the square loss and a linear hypothesis space, the regularized empirical risk is given by

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

where $\lambda > 0$. [For convenience, we've dropped the $\frac{1}{n}$ scale factor from the empirical risk term.]

1. [2] **Show that** for w to be a minimizer of $J(w)$, we must have $X^T Xw + \lambda I w = X^T y$, and **give an expression** for the minimizer. **Justify** that the expression is valid (e.g. if you invert a matrix, justify that it's invertible.), and **note any issues** we would have if we took $\lambda = 0$ (e.g. what happens to the solution if X is a matrix of zeros?).

   SOLUTION: First order condition is

$$\begin{aligned} & & \partial_w \left[(Xw - y)^T (Xw - y) + \lambda w^T I w \right] \\ & = & 2X^T Xw - 2X^T y + 2\lambda w = 0 \\ \iff X^T Xw + \lambda I w & = & X^T y \end{aligned}$$

   Solving for w, the minimizer is

$$w = \left(X^T X + \lambda I\right)^{-1} X^T y.$$

   The inverse always exists when $\lambda > 0$ because $X^T X$ is symmetric positive semidefinite, and when we add λI (a symmetric positive definite matrix), we get a symmetric positive definite matrix, which is invertible. If $\lambda = 0$, we get $\left(X^T X\right)^{-1}$, but $X^T X$ may not be invertible.

4 Perceptron

The **perceptron loss** is given by

$$\ell(\hat{y}, y) = \max\{0, -\hat{y}y\}.$$

And consider the hypothesis space of linear functions $\mathcal{H} = \left\{f \mid f(x) = w^T x,\, w \in \mathbf{R}^d\right\}$.

1. [1] Is the perceptron loss a margin loss? Justify your answer.
SOLUTION: A margin loss is a loss function that depends on y and $\hat{y}$ only via the "margin", which is the product $y\hat{y}$. ℓ is clearly a margin loss.

2. [1] Suppose we have a linear function $f(x) = w^T x$, for some $w \in \mathbf{R}^d$. Geometrically, we say that the hyperplane $H = \{x \mid f(x) = 0\}$ separates the dataset $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n)) \in \mathbf{R}^d \times \{-1, 1\}$ if all x_i corresponding to $y_i = -1$ are strictly on one side of H, and all x_i corresponding to $y_i = 1$ are strictly on the other side of H. ("Strictly" here means that no x_i's lie on H.) Give a mathematical formulation of the necessary and sufficient conditions for $f(x) = w^T x$ to separate $\mathcal{D}$. [Hint: Answer will involve the data points and the function f.]
SOLUTION:
$$y_i f(x_i) > 0 \; \forall i \in \{1, \ldots, n\}$$

3. [1] In the homework we showed that if our prediction function $f(x) = w^T x$ separates a dataset $\mathcal{D}$, then the total perceptron loss on $\mathcal{D}$ is 0. The converse is not true: we may have total perceptron loss 0, but $f(x)$ may not separate $\mathcal{D}$. Explain how this can happen.
SOLUTION: We may have this if any x_i lies on the hyperplane — i.e. if $f(x_i) = w^T x_i = 0$. When this happens, the loss on this example will be 0, but the point is not strictly on the correct side of the hyperplane.

5 Regularized Perceptron

Consider a hypothesis space of linear functions $\mathcal{H} = \left\{f \mid f(x) = w^T x,\ w \in \mathbf{R}^d\right\}$. Let $\ell(\hat{y}, y) = \max\{0, -\hat{y}y\}$ be the Perceptron loss. Consider the objective function

$$J(w) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left\{0, -y_i w^T x_i\right\}.$$

We are interested in finding the minimizer of $J(w)$ **subject to** the constraint that $\|w\|^2 \geq 1$.

1. [2] Let $J_1(w; x, y) = \frac{1}{2}\|w\|^2 + c\max\left\{0, -yw^T x\right\}$. Give a subgradient g of $J_1(w; x, y)$ with respect to w. The subgradient will be a function of x, y, c, and w.
SOLUTION:

$$g = \begin{cases} -cyx + w & \text{for } yw^T x < 0 \\ w & \text{for } yw^T x \geq 0. \end{cases}$$

2. [1] Write the Lagrangian for the problem of minimizing $J(w)$ **subject to** the constraint that $\|w\|^2 \geq 1$.
SOLUTION:

$$L(w, \lambda) = J(w) + \lambda\left(1 - \|w\|^2\right)$$

3. [2] Assuming it's attained, give an expression for the [primal] optimal value of the optimization problem in terms of the Lagrangian. **Explain** why this gives the same optimal value as the original problem.
SOLUTION: The primal optimal value is

$$p^* = \min_{w} \sup_{\lambda \geq 0} L(w, \lambda)$$

for the following reason: If w is feasible, then the inner supremum is just $J(w)$, and otherise it's ∞. The outer minimum will only ever select w for which the inner optimization is $J(w)$. So it's equivalent to the original problem.

4. [1] State the dual objective function and the dual optimization problem in terms of the Lagrangian function.
SOLUTION: The dual objective function is

$$g(\lambda) = \inf_{w} L(w, \lambda),$$

and the dual optimization problem is

$$\sup_{\lambda \geq 0} \inf_{w} L(w, \lambda)$$

5. [1] $J(w)$ is not differentiable. Give an equivalent optimization problem that has a differentiable objective function. [Hint: You may want to introduce new variables as we did for the SVM.][NOTE: This problem should also have required that the constraint functions be either linear or quadratic.)]
SOLUTION:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{such that} \quad & \xi_i \geq 0\ \forall i \\ & \xi_i \geq -y_i w^T x_i \leq 0\ \forall i \\ & w^T w \geq 1 \end{aligned}$$

This would be sufficient. But we can also put it into standard form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{such that} \quad & -\xi_i \leq 0\ \forall i \\ & -\xi_i - y_i w^T x_i \leq 0\ \forall i \\ & 1 - w^T w \leq 0 \end{aligned}$$

6. [1] Is this a convex optimization problem? Why or why not?
SOLUTION: This is not a convex optimization problem. A convex optimization problem must have a convex feasible set. [Recall: The feasible set is the set that satisfies all the constraints.] The set of w satisfiying $\|w\|^2 \geq 1$ is not a convex set. You can also see this from the standard form in the previous problem. The function $1 - w^T w$ is concave, not convex.

7. [1] There's a good reason for the constraint $\|w\|^2 \geq 1$. What is the **unconstrained** minimizer of $J(w)$? Explain your answer. [Hint: This does not require calculation.]
SOLUTION: At $w = 0$, the objective function is 0. Since the objective function is always nonnegative, $w = 0$ is a solution to the unconstrained problem.

6 Regularization and the Bias Term

Suppose our input space is $\mathcal{X} = \mathbf{R}^2$ and our output space is $\mathcal{Y} = \mathbf{R}$. We have n labeled training points that we put together into a design matrix $X \in \mathbf{R}^{n \times 2}$. Let $y \in \mathbf{R}^n$ represent the vector of outputs. **This data stays fixed for all parts of this problem**. We'd like to find an affine function that fits this data. Let's create an "augmented design matrix" matrix $X_b \in \mathbf{R}^{n \times 3}$, whose first column is $b = (B, \ldots, B) \in \mathbf{R}^{n \times 1}$, for some $B > 0$. The next two columns are the original matrix X. Suppose X_b is full rank, and let

$$w^* = \underset{w \in \mathbf{R}^3}{\arg\min} \|X_b w - y\|^2.$$

Let's write the components of w^* as $w^* = (w_0, w_1, w_2)$.

1. [1] Given a new $x = (x_1, x_2)$, give an expression for the prediction on x corresponding to w^*. (Hint: This expression should be in terms of w_0, w_1, w_2, x_1, x_2 and B).
   SOLUTION:
   $$x \mapsto w_0 B + w_1 x_1 + w_2 x_2$$

2. [1] Let w^1 be the solution with $B = 1$ and let w^{100} be the solution with $B = 100$. If $f_1(x)$ is the prediction corresponding to w^1 and $f_{100}(x)$ is the prediction corresponding to w^{100}, then for all x we have (choose one of the following): [Below, w_0^{100} refers to the first component of w^{100}, etc.]

   (a) $f_1(x) < f_{100}(x)$
   (b) $f_1(x) > f_{100}(x)$
   (c) $w_0^1 = w_0^{100}$and $f_1(x) = f_{100}(x)$
   (d) $w_0^1 < w_0^{100}$ and $f_1(x) = f_{100}(x)$
   (e) $w_0^1 > w_0^{100}$ and $f_1(x) = f_{100}(x)$

   SOLUTION: (e) The hypothesis space is the same regardless of the value of B, so without regularization, we will get $f_1(x) = f_{100}(x)$. Without regularization, we will have $w_0^{100} = \frac{1}{100} w_0^1$.

3. [1] Suppose for $B = 1$, the prediction function is $x \mapsto 5.3 + 2.0x_1 + 0.9x_2$. When we introduce an ℓ_2 regularization term, the resulting prediction function is $x \mapsto 1.99 + 2.10x_1 + 1.06x_2$. Now suppose we refit the model with ℓ_2 regularization but with $B = 100$. Which of the following prediction functions seems most likely to be the result (NOTE: these functions are coming from an iterative optimization algorithm and may have some optimization error):

   (a) $x \mapsto 1.99 + 2.10x_1 + 1.06x_2$
   (b) $x \mapsto 0.68 + 2.15x_1 + 1.13x_2$
   (c) $x \mapsto 5.49 + 1.96x_1 + 0.88x_2$
   SOLUTION: (c) There are 3 options: the bias term stays the same, gets bigger, or goes smaller. Without regularization, $w_0^{100} = \frac{1}{100} w_0^1$. So when $B = 100$, we expect the coefficient of B to be much smaller than when $B = 1$. Thus it decreases the amount of ℓ_2 penalty, which means we expect it to be closer to its unregularized value, which corresponds to a bias term of ~ 5.3.

4. [1] Suppose $B = 1$. Let's introduce a new feature that is a duplicate of x_2, namely $x_3 = x_2$. Suppose we fit the model with ℓ_2 regularization. Which of the following prediction functions seems most likely to result? (NOTE: these functions are coming from an iterative optimization algorithm and may have some optimization error):

   (a) $x \mapsto 1.98 + 2.09x_1 + 0.20x_2 + 0.88x_3$
   (b) $x \mapsto 1.98 + 2.09x_1 + 0.54x_2 + 0.54x_3$
   (c) Both seem equally likely
   SOLUTION: (b). Both (a) and (b) give the same predictions, and thus have the same empirical loss. However, the ℓ_2 penalty for (b) is smaller, because $.54^2 + .54^2 < .2^2 + .88^2$. You don't need to calculate to know this: in general, if we minimize $a_1^2 + \cdots + a_d^2$ subject to $a_1 + \cdots + a_d = c$, the solution is $a_i = c/d$, for all $i = 1, \ldots, d$. You should know this result.

5. [1] Suppose $B = 1$. With **lasso** ℓ_1 regularization and the original feature set, the prediction function is $x \mapsto 4.47 + 2.05x_1 + 0.93x_2$. Let's introduce a duplicate feature $x_3 = x_2$. Suppose we refit with the same ℓ_1 regularization penalty. Which of the following prediction functions seems most likely to result? (NOTE: these functions are coming from an iterative optimization algorithm and may have some optimization error):

   (a) $x \mapsto 4.45 + 2.05x_1 + 0x_2 + 0.93x_3$

   (b) $x \mapsto 4.45 + 2.05x_1 + .93x_2 + 0x_3$

   (c) $x \mapsto 4.45 + 2.05x_1 + .46x_2 + .47x_3$

   (d) They are equally plausible results.
   SOLUTION: (d) Note that (a), (b), and (c) all give the same predictions, and all have the same ℓ_1 regularization term. Thus they are equivalent minimizers of the objective function. Lasso gives sparsity sometimes, but not always. When we have duplicate features, lasso doesn't care how the weight gets divided amont the duplicate features.

6. [1] Suppose $B = 1$. With **lasso** ℓ_1 regularization and the original feature set, the prediction function is $x \mapsto 4.47 + 2.05x_1 + 0.93x_2$. Let's introduce a new feature that is a **multiple** of x_2, namely $x_3 = 2x_2$. Suppose we refit with the same ℓ_1 regularization penalty. Which of the following prediction functions seems most likely to result? (NOTE: these functions are coming from an iterative optimization algorithm and may have some optimization error):

   (a) $x \mapsto 4.45 + 2.05x_1 + 0x_2 + 0.47x_3$

   (b) $x \mapsto 4.45 + 2.05x_1 + .94x_2 + 0x_3$

   (c) $x \mapsto 4.45 + 2.05x_1 + .71x_2 + .11x_3$

   (d) $x \mapsto 4.45 + 2.05x_1 + .34x_2 + .30x_3$

   (e) All seem equally likely
   SOLUTION: (a). All 4 gives the same predictions, but (a) has the smallest ℓ_1 regularization penalty.