Machine Learning and Computational Statistics, Spring 2015
Homework 6: Midterm Review

Due: Friday April 10, 2015, at 4pm (Submit via NYU Classes)

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. You may include your code inline or submit it as a separate file. You may either scan hand-written work or, preferably, write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or MathJax via iPython).

1 Robust Ridge Regression

1. Lasso regression uses a square loss with ℓ_1 regularization. Here we consider an ℓ_1 loss with ℓ_2 regularization. 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 $y_i \in \mathcal{Y} = \mathbf{R}$, for all $i = 1, \ldots, n$. Write the objective function for ℓ_2-regularized emprical risk minimization with an ℓ_1 loss, over a linear hypothesis space.

2. Show that the objective function is convex. (You are free to cite any of the facts given in Appendix A).

3. Write the objective function as a quadratic program (i.e. an optimization problem with a quadratic objective and linear constraints).

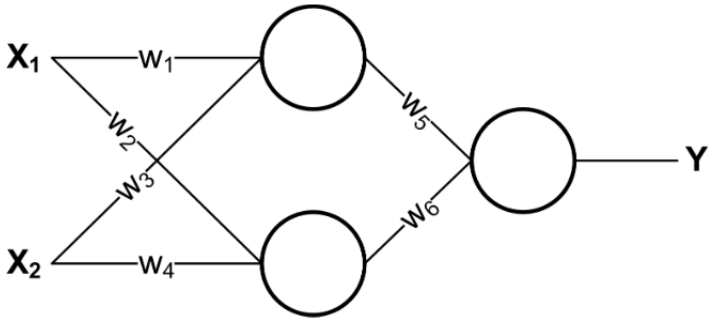
2 RBF Kernel

1. The RBF kernel $k(x_i, x_j) = \exp\left(-\frac{1}{2}\|x_i - x_j\|^2\right)$ is symmetric, positive semidefinite, and thus by Mercer's theorem we know there exists an inner product space $\mathcal{H}$ (i.e. a space with an inner product $\langle \cdot, \cdot \rangle$ defined) and a "feature mapping" $\phi : \mathbf{R}^d \to \mathcal{H}$ such that $k(x_i,x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. Show that the distance between the feature representations of any two points x_i and x_j in the space $\mathcal{H}$ is at most $\sqrt{2}$. (Hint: In an inner product space, the distance between two elements x and y is defined as $\|x - y\|$, and $\|x\| = \sqrt{\langle x, x \rangle}$. Also, See Homeork #4 Problem 2.1.)

3 Regularized Logistic Regression

(Source: Jaakkloa.) Consider minimizing

$$J(w) = -L(w, \mathcal{D}_{\text{train}}) + \lambda \|w\|^2$$



where

$$L(w, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \log \sigma \left(y_i x_i^T w \right),$$

and where $\sigma(a) = 1/\left(1 + e^{-a}\right)$, denotes the size of the set $\mathcal{D}$, and $y_i \in \{-1, 1\}$. Answer the following true/false questions with brief explanation (so it's clear you're not just guessing):

1. $J(w)$ has multiple locally optimal solutions? T/F?
2. Let $\hat{w} = \arg\min_w J(w)$ be a global optimum. Then $\hat{w}$ is sparse (i.e. has many zero entries). T/F?
3. [Optional] If the training data are linearly separable, then some weights w_j might become infinite if $\lambda = 0$. T/F?
4. $L(\hat{w}, \mathcal{D}_{\text{train}})$ always increases as we increase λ. [NOTE: L is the log-likelihood, and the negative empirical risk.] T/F?
5. $L(\hat{w}, \mathcal{D}_{\text{test}})$ always increases as we increase λ. T/F?

4 Neural Networks with Linear Activation Function

Suppose we have the neural network shown below with linear activation function. That is, the output of the top hidden node is $h_1 = c\,(x_1 w_1 + x_2 w_3)$ and the bottom hidden node outputs $h_2 = c(x_1 w_2 + x_2 w_4)$. The output node is just $Y = h_1 w_5 + h_2 w_6$.

1. Redesign the neural network to compute the same function without using any hidden units. Draw the equivalent network and give expressions detailing for the new weights in terms of the old weights and the constant c.
2. Can the space of functions that is represented by the above neural network also be represented by linear regression?

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 $\mathbf{R}^n$
- Log-Sum-Exp[1]: $(x_1, \ldots, x_n) \mapsto \log\left(e^{x_1} + \cdots + e^{x_n}\right)$ 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. More generally, if $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, and if $w(y) \geq 0$ for each $y \in \mathcal{A}$, then the function

$$g(x) = \int_{\mathcal{A}} w(y) f(x, y)\, dy$$

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

$$g(x) = f\left(Ax + b\right),$$

with $\mathbf{dom}\ g = \{x \mid Ax + b \in \mathbf{dom}\ f\}$. 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.

[1] This function can be interpreted as a differentiable (in fact, analytic) approximation to the max function, since

$$\max\{x_1, \ldots, x_n\} \leq \log\left(e^{x_1} + \cdots + e^{x_n}\right) \leq \max\{x_1, \ldots, x_n\} + \log n.$$

Can you prove it? Hint: $\max(a, b) \leq a + b \leq 2\max(a, b)$.

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 $\mathbf{dom}\ f = \mathbf{dom}\ f_1 \cap \cdots \cap \mathbf{dom}\ f_m$.

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