Homework 2
10-702/36-702 Statistical Machine Learning

Due: Friday Feb 4 3:00

Hand in to: Michelle Martin GHC 8001.

1 Convexity and Optimization

1. (Convexity)

   (a) Show that $1/g(x)$ is convex if g is twice-differentiable, concave and positive (hint: use composition property).

   (b) We know that for every point on the boundary of a convex set, there exists a supporting hyperplane. The converse is not necessarily true. Give an example of a non-convex set for which there exists a supporting hyperplane for every point on its boundary.

2. (Subdifferentials) For Lasso problems, we want to solve an optimization of $\arg\min_{\beta\in\mathbb{R}^n} \frac{1}{2}||X\beta - Y|| + \lambda||\beta||_1$ where X is $m \times n$ design matrix and Y is a $m \times 1$ response vector.

   In this problem, we will assume design matrix is identity and solve the simpler problem of $\arg\min_{z\in\mathbb{R}^n} \frac{1}{2}||z - y||_2^2 + \lambda||z||_1$ where y is the $n \times 1$ response vector. Gradient descent algorithm for actual Lasso solves this simpler problem many times as intermediate steps.

   Provide a proof for all your solutions:

   (a) Let $f : \mathbb{R} \to \mathbb{R}$, find the subdifferential of $f(z) = |z|$

   (b) Let $f : \mathbb{R}^n \to \mathbb{R}$, find the subdifferential of $f(z) = ||z||_1$

   (c) Let $f : \mathbb{R}^n \to \mathbb{R}$, let $y \in \mathbb{R}^n$ be a fixed vector, find the subdifferential of $f(z) = \frac{1}{2}||z - y||_2^2 + \lambda||z||_1$

   (d) Using the fact that z minimizes f if and only if $0 \in \partial f(z)$, prove that we can find z^*, the minimizer of $\frac{1}{2}||z - y||_2^2 + \lambda||z||_1$, by soft-thresholding y, i.e.

$$z^*(i) = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \\ 0 & \text{if } -\lambda \le y_i \le \lambda \end{cases}$$

3. (Optimization) In class, we considered the following entropy maximization optimization problem:

$$\min \sum_{i=1}^{n} x_i \log x_i$$
$$s.t.\ Ax \le b$$
$$1^T x = 1$$

   (a) Argue that the primal is a convex optimization problem.

   (b) Write down the KKT conditions for this problem.

   (c) Using the KKT conditions, show that the primal optimal solution x^* (the maximum entropy distribution under given constraints) can be obtained if you are given a dual optimal solution λ^*.

2 Density Estimation

Let $X_1, \ldots, X_n$ be a sample from a distribution P with density p where $X_i \in [0,1]^d$. Divide the cube $[0,1]^d$ into sub-cubes $B_1, \ldots, B_N$ with sides of length h. The number of sub-cubes is thus $N = (1/h)^d$. The histogram density estimator is

$$\widehat{p}_h(x) = \sum_{j=1}^{N} \frac{\widehat{\pi}_j}{h^d} I(x \in B_j)$$

where $\widehat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in B_j)$. Assume that $p \in \mathcal{P}$ where

$$\mathcal{P} = \left\{ p : \ |p(x) - p(y)| \le L\,||x - y|| \right\}.$$

(a) Find an upper bound on $R(h) = \mathbb{E}|\widehat{p}_h(x) - p(x)|^2$. (Hint: find an upper bound on the bias and variance of $\widehat{p}_h(x)$.)

(b) Find h_n to minimize (the bound on) $R(h)$. At what rate does $R(h_n)$ go to 0?

(c) Derive an exponential inequality for $|\widehat{p}_h(x) - p(x)|$.

(d) Prove a concentration inequality for $\int |p(x) - \widehat{p}(x)|dx$. The following result will be helpful: If $(Y_1, \ldots, Y_k)$ is a random vector with a Multinomial$(n, p_1, \ldots, p_k)$ distribution and if $0 < \epsilon < 1$ and $k/n \le \epsilon^2/20$ then

$$\mathbb{P}\left(\sum_{j=1}^{k} |Y_j - \mathbb{E}(Y_j)| \ge n\epsilon \right) \le 3e^{-n\epsilon^2/25}.$$