

Statistical Learning, Spring 2013

Homework exercise 1

Due date: 9 April 2013

1. **Population Optimizer of absolute loss**

Prove that for absolute loss: $L_{\text{abs}}(Y, f(X)) = |Y - f(X)|$, EPE is minimized by setting $f^*(x) = \text{Median}(Y|X = x)$

Hint: you may find the following identity useful:

$$\int_{y>c} (y - c)dP(y) = \int_{y>c} Pr(Y > y)dy$$

* **Extra credit: generalization to quantile loss** The τ th quantile loss for $0 < \tau < 1$ is defined as:

$$L_{\tau}(Y, f(X)) = \begin{cases} \tau \times (Y - f(X)) & \text{if } Y - f(X) > 0 \\ -(1 - \tau) \times (Y - f(X)) & \text{otherwise} \end{cases}$$

Prove that the EPE is minimized by setting $f^*(x)$ to be the τ th quantile of $P(Y|X = x)$, i.e., $P(Y \leq f^*(x)|X = x) = \tau$

2. **ESL 2.3:** Derive equation (2.24) (expected median distance to origin's nearest neighbor in an ℓ_p ball):

$$d(p, n) = \left(1 - \frac{1}{2}\right)^{1/p}$$

Suggested approach:

- (a) Find the probability that all observations are outside a ball of radius $r < 1$, as a function of r .
- (b) You are looking for r such that this probability is $1/2$.

Plot $d(p, n)$ against p for $n \in \{100, 5000, 100000\}$ and $p \in \{3, 5, 10, 20, 50, 100\}$ (make one curve for every value of n — use the R functions `plot()` and `lines()`) and interpret the graph.

3. **ESL 2.7:** Compare classification performance of k-NN and linear regression on the `zipcode`¹ data, on the task of separating the digits 2 and 3. Use $k \in \{1, 3, 5, 7, 15\}$. Plot training and test error for k-NN choices and linear regression. Comment on the shape of the graph.

4. **Guaranteed error reduction via Ridge Regression**

Assume the linear model is correct, i.e., $E(Y|X = x) = x^T \beta$. Consider making a prediction at a new point x_0 based on a Ridge Regression with smoothing parameter λ : $\hat{Y} = x_0^T \hat{\beta}^{\text{ridge}}(\lambda)$

- (a) Derive explicit expressions for the bias and variance of \hat{Y} as a function of λ (use the SVD of X for the variance).
- (b) Set $MSE(\lambda) = \text{bias}^2(\lambda) + \text{Var}(\lambda)$ from above, show that

$$\left. \frac{d}{d\lambda} MSE(\lambda) \right|_{\lambda=0} < 0$$

Suggested approach:

- i. Show by differentiation that $\left. \frac{d}{d\lambda} \text{Var}(\lambda) \right|_{\lambda=0} < 0$.
- ii. Show that $\left. \frac{d}{d\lambda} \text{bias}^2(\lambda) \right|_{\lambda=0} = 0$. Look at the expression for bias to find a simple argument, avoid complex differentiations!
- (c) Briefly explain the meaning of this result — what happens when we add *a little* ridge penalty to standard linear regression?

Surprisingly, the same is true for the Lasso. The proof, however, is much more involved.

* **Extra credit problem: Optimality of k-NN in fixed dimension**

Assume $X \sim \text{Unif}([0, 1]^p)$, and $Y = f(X) + \epsilon$ with $\epsilon \sim (0, \sigma^2)$ (that is, $f(x) = E(Y|X = x)$).

Assume f is Lipschitz: $\|x_1 - x_2\| < \delta \Rightarrow |f(x_1) - f(x_2)| < c\delta$, $\forall x_1, x_2 \in [0, 1]^p$. Choose any sequence $k(n)$ such that:

$$\begin{aligned} k(n) &\xrightarrow{n \rightarrow \infty} \infty \\ k(n)/n &\xrightarrow{n \rightarrow \infty} 0 \end{aligned}$$

Then:

$$\text{EPE}(\text{k-NN using } k(n)) \xrightarrow{n \rightarrow \infty} \text{EPE}(f) = \sigma^2$$

(The proof does not have to be completely formal, for example you can replace a binomial with its normal approximation without proof of the relevant asymptotics).

¹Training: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/zip.train.gz>
 Testing: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/zip.test.gz>
 Info: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/zip.info>