

STAT-598Y HW 1

DUE ON 9/10/09 IN CLASS

Problem. 1. Consider the binary classification problem where $\mathcal{Y} = \{0, 1\}$. Suppose we use the loss function $\ell(y, p)$ defined as follows (y is true label, p is your prediction):

$$\ell(y, p) = \begin{cases} 0, & y = p \\ 2, & y = 1, p = 0 \\ 1, & y = 0, p = 1. \end{cases}$$

In other words, the cost of mis-classify “1” to “0” is twice as large as the cost of mis-classify “0” to “1”. Identify the optimal classifier $h^*(x)$ and prove that it indeed achieves the minimum risk, i.e. $R^* = R(h^*)$.

Problem. 2. Consider the simplest “regression” problem where we only observe the response variables. Another way to think this is that you have $x_1 = \dots = x_n = 1$, i.e. every x is always the constant 1. Suppose $Y \sim F(\mu, \sigma^2)$ with $\mathbb{E}[Y] = \mu$, $\mathbb{V}[Y] = \sigma^2 < \infty$. Furthermore, assume that the distribution of Y is **symmetric** around its mean μ . We will use the expected risk $R_1(\hat{\theta}) = \mathbb{E}_{\mathcal{D}_n}[|\hat{\theta} - \mu|]$ or $R_2(\hat{\theta}) = \mathbb{E}_{\mathcal{D}_n}[(\hat{\theta} - \mu)^2]$ to evaluate the performance of an estimator $\hat{\theta}$. (You can use either R or matlab to do the computation.)

(1) Show that the the empirical risk minimizer obtained using the absolute error loss function $\hat{\theta}_1 = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n |y_i - \theta|$ is just the sample median, and the empirical risk minimizer of μ obtained using squared error loss function $\hat{\theta}_2 = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n (y_i - \theta)^2$ is just the sample mean.

(2) Suppose Y is normally distributed with $\mu = 0$ and $\sigma^2 = 1$. Compare $R_1(\hat{\theta}_1)$ vs $R_1(\hat{\theta}_2)$ and $R_2(\hat{\theta}_1)$ vs $R_2(\hat{\theta}_2)$ for $n = 50, 200, 1000$ training samples. *Hint: to compute the expected risk you can repeat the experiment 100 times and take the average as an approximate to R_1 (or R_2).*

(3) Suppose Y has a double-exponential distribution (aka Laplace) with pdf $f(y) = \exp(-|y|)/2$. Compare $R_1(\hat{\theta}_1)$ vs $R_1(\hat{\theta}_2)$ and $R_2(\hat{\theta}_1)$ vs $R_2(\hat{\theta}_2)$ for $n = 50, 200, 1000$ training samples.

What conclusions can you draw about the loss function used in evaluating risk and the loss function in computing the empirical risk minimizer? Try some other symmetric distributions for Y to confirm.

Problem. 3. (1) Prove Theorem 1-2 in lecture notes. (2) Suppose for regression problems we use the absolute loss function $\ell(y, p) = |y - p|$. Prove that the optimal risk is achieved by the conditional median function, i.e. $h^*(x)$ satisfies for any x , $P(Y \leq h^*(x) | X = x) = 1/2$.