

Lecture notes for ECE 695- 09/08/03.

1 Label Prediction: Problem setup

Unknown distribution P over $X \times \{0, 1\}$.

Input: $S \in \{X \times \{0, 1\}\}^m$, S is an i.i.d sample from P . $S = \{(x_i, y_i) : 1 \leq i \leq m\}$ is called Training data.

Output: $h : X \rightarrow \{0, 1\}$. h called hypothesis or predictor.

Cost: $E^P(h) = Pr(\{(x, y) : h(x) \neq y\})$. $E^P(h)$ called true error.

The problem is that we have to find a "good" h . A naive approach will encounter overfitting.

Basic solution: Limit the search to some predesignated set H called the hypothesis class.

First theoretical task: Provide guarantees for a good selection of H that avoids overfitting.

First, we use basic probability considerations to give guarantees for selections from H that overfit. Fix some h and assume that $E^P(h) > \epsilon$. The probability that $E^S(h) = 0$ is at most $(1 - \epsilon)^{|S|}$. So, we can conclude that $\forall P, \forall H, \forall m, \forall \delta$ with probability $> 1 - \delta$ if some $h \in H$ has $E^S(h) = 0$ then,

$$E^P(h) < \frac{1}{m}(\ln(|H|) + \ln(\frac{1}{\delta})), \quad (1)$$

where $m = |S|$.

Equivalently, $\forall P, \forall H, \forall \epsilon, \forall \delta$, with probability $> 1 - \delta$ if some $h \in H$ has $E^S(h) = 0$ and

$$m > \frac{1}{\epsilon}(\ln(|H|) + \ln(\frac{1}{\delta})), \quad (2)$$

then $E^P(h) < \epsilon$ where $m = |S|$.

We wish to analyze the error of h 's that do not necessarily have zero empirical error. The main tool that will be used is the Chernoff/Hoeffding bound.

Chernoff/Hoeffding Bound: Let $\{X_i\}_{i=1}^{\infty}$ be independent Bernoulli random variables each with expectation $0 \leq p \leq 1$. Then, $\forall m$

$$Pr\left(\left|\frac{1}{m}\sum_{i=1}^m X_i - p\right| > \epsilon\right) < 2e^{-2m\epsilon^2}.$$

In our case, we use it as follows: Fix any $h : X \rightarrow \{0, 1\}$ and let

$$X_i = \begin{cases} 1 & \text{if } h(x_i) \neq y_i \\ 0 & \text{otherwise.} \end{cases}$$

Note that $E(X_i) = E^P(h)$. By Hoeffding's inequality,

$$Pr(|E^S(h) - E^P(h)| > \epsilon) < 2e^{-2m\epsilon^2}.$$

It can also be shown that

$$Pr(E^P(h) - E^S(h) > \epsilon) < e^{-2m\epsilon^2}. \quad (3)$$

Exercise: Prove that $\forall P, \forall H, \forall m, \forall \delta$, with probability $> 1 - \delta$, for all $h \in H$

$$E^P(h) - E^S(h) < \sqrt{\frac{(\ln(|H|) + \ln(1/\delta))}{2m}}, \quad (4)$$

where $m = |S|$.

Bounds depending on description length: Given H , we fix a description of the hypothesis in H as follows. Let $(\hat{\cdot}) : H \rightarrow \{0, 1\}^*$ map all $h \in H$ to binary strings of finite length. Thus, \hat{h} is the string coding h and let $|\hat{h}|$ denote the description length of h . Assume that the above binary mapping is prefix free. Then, by Kraft's inequality,

$$\sum_{h \in H} 2^{-|\hat{h}|} \leq 1. \quad (5)$$

Suppose that we have different error tolerances for different h viz., let ϵ_h be the error tolerance for h . Set

$$\epsilon_h = \sqrt{\frac{(\ln 2(|\hat{h}|) + \ln(1/\delta))}{2m}},$$

and by (3) we get for every fixed h ,

$$Pr(E^P(h) - E^S(h) > \epsilon_h) < \frac{\delta}{2^{|\hat{h}|}}$$

Using the union bound and Kraft inequality (5), we can conclude that

$$Pr(\exists h \in H : E^P(h) - E^S(h) > \epsilon_h) < \delta,$$

OR

With probability atleast $1 - \delta$, $\forall h \in H$, $E^P(h) - E^S(h) \leq \epsilon_h$.

More generally, we can say that $\forall P, \forall H, \forall m, \forall \delta$, with probability $> 1 - \delta$ for all $h \in H$

$$E^P(h) - E^S(h) < \sqrt{\frac{(\ln 2)(|\hat{h}|) + \ln(1/\delta)}{2m}}, \quad (6)$$

where $m = |S|$.

Bayesian Approach: We model our prior knowledge by a prior distribution over the hypotheses. This is equivalent to specifying a particular description of H . For example, for a 2-adic distribution $q(h)$ over H we can choose binary Huffman coding as the description mapping. Then $q(h) = 2^{-|\hat{h}|}$ and

$$E^P(h) - E^S(h) < \sqrt{\frac{(\ln(1/q(h)) + \ln(1/\delta))}{2m}},$$

Relations between different predictors Note that there exists some $f : X \rightarrow \{0, 1\}$ minimizing $E^P(\cdot)$. It is given by

$$f(x) = \begin{cases} 1 & \text{if } \frac{P(x,1)}{P(x,0)+P(x,1)} > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Ofcourse since P is unknown, f is unknown. So, let us constrain ourselves to a class H of predictors. Let h^{**} be the predictor that minimizes $E^S(\cdot)$ and h^* be the predictor that minimizes $E^P(\cdot)$ over H . By definition,

$$E^S(h^{**}) \leq E^S(h^*), \quad (8)$$

$$E^P(h^*) \leq E^P(h^{**}). \quad (9)$$

Then it follows from our arguments that

$$E^P(h^{**}) < E^S(h^{**}) + \epsilon \leq E^S(h^*) + \epsilon \leq E^P(h^*) + 2\epsilon.$$

This gives us a handle on the difference between the true errors of the samplewise optimal predictor and the best predictor (in terms of minimizing the true error).