CSE 254: Statistical Machine Learning **Winter 2011**

Topic 4 — On the Generalization Ability of On-Line Learning Algorithms

4.1 The problem

Online algorithms, such as the Perceptron algorithm, hava a guaranteed bound on the cumulative loss on the sequence itself. These bounds hold for *every* sequence. The Perceptron algorithm makes at most $(R/\gamma)^2$ mistakes where R is the radius of the ball containing the examples and γ is the classification margin.

We define a "comparison class" $\mathcal{H}$ of prediction functions, and a loss function ℓ. The loss of $h \in \mathcal{H}$ on the example (X, Y) is $0 \leq \ell(h(X), Y) \leq 1$. Given a training set $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn IID from some fixed distribution $\mathcal{D}$, find $\widehat{H}$, whose generalization error is not much worse than the best rule in $\mathcal{H}$.

$$\text{risk}(h) = E\ell(h(X), Y)$$

$$P\left(\text{risk}(\hat{H}) \geq \inf_{h \in \mathcal{H}} \text{risk} h + \epsilon\right) \leq \delta$$

4.2 Simple Solutions

Uniform convergence method, minimize

$$\text{risk}_{\text{emp}}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$$

Given that the VC dimension of $\mathcal{H}$ is d then we have the bound

$$\text{risk}(\hat{H}) \leq \text{risk}(h^*) + 2c\sqrt{\frac{d + \ln(2/\delta)}{n}}$$

Plug in a random place algorithm. The total loss of the online algoirthm is

$$M_n = \frac{1}{n} \sum_{t=1}^{n} \ell(H_{t-1}(X_t), Y_t)$$

By considering the average over all n hypotheses we get:

$$P\left(\frac{1}{n} \sum_{t=1}^{n} \text{risk}(H_{t-1}) \geq M_n + \sqrt{\frac{2}{n} \ln \frac{1}{\delta}}\right) \leq \delta$$

4.3 Test-on-Rest solution

4.4 Application to SVM