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

Topic 4 — Uniform Convergence Bounds

4.1 Occam's Razor

Let H be a finite set of classifiers, let the reqired reliability be $\delta > 0$ and the number of training examples be m, then with probability at least $1-\delta$ over the random choice of the training set, for all $h \in H$:

$$\mathrm{err}(h) \le \widehat{\mathrm{err}}(h) + \sqrt{\frac{\ln 2|H| + \ln(1/\delta)}{2m}}$$

In particular this holds for the rule that minimizes the emprical error.

4.2 Dichotomies

When the concept class is infinite we need to measure it's complexity in a different way. The set of dichotomies for a set of instances $S = \{x_1, x_2, \ldots, x_m\}$ is

$$\Pi_H(S) = \{< h(x_1), h(x_2), \ldots, h(x_m) >: h \in H\}$$

The size of $\Pi_H(S)$ is at most 2^m, in which case we say that S is shattered.

The growth function $\Pi_H(m)$ is equal to the maximal size of $\Pi_H(S)$ for all sets S of size m.

4.3 Bound using growth function

Let H be a finite set of classifiers, whose growth function is $\Pi_H(m)$. Let the reqired reliability be $\delta > 0$ and the number of training examples be m, then with probability at least $1-\delta$ over the random choice of the training set, for all $h \in H$:

$$\mathrm{err}(h) \le \widehat{\mathrm{err}}(h) + \sqrt{\frac{32(\ln \Pi_H(m) + \ln(8/\delta)}{m}}$$

Consider Glivenko-Canteli and then the general case.

We are bounding the difference between the emprirical and the true distributions of a class of sets.

Notation: Z_i are the IID drawn samples. The true probability of A is $\nu(A) = P(Z_1 \in A)$, the empirical probability of A is the random variable $\nu_n(A) = (1/n)\sum_{j=1}^{n} I_{Z_j \in A}$

1. **ghost sample** The first symmetrization step is to take a second "ghost" sample and consider the differences between the empirical probabilities of the sets in the two samples.

   Basic idea: consider the set A^* which maximizes $|\nu_n(A) - \nu(A)|$. Conditioned on this gap being larger than ϵ we still have that with probability at least $(1/2)$ $|\nu'_n(A^*) - \nu_n(A^*)| < \epsilon/2$

2. **random signs** The second step is to put random n signs. We can think of it as determining for each pair of example Z_i, Z_i' whether they stay in their respective sets or whether they are switched. This clearly does not change the probability of any condition because the two configurations have the same probability.

   We can now remove the ghost sample by considering that if $|A| \leq \epsilon/4$ and $|B| \leq \epsilon/4$ then $|A-B| \leq \epsilon/2$.

3. **Conditioning** The third step is to condition on the first sample $Z_1', \ldots, Z_n'$ and consider all possible labelings that can be produced on this set.

   We bound the probability of samples that yield a large difference between the two samples separately for each labeling and use the union bound over the labelings.

4. **Hoeffding** We use Hoeffding to bound the sum for each sampe separately.

4.4 Sauer's Lemma

VC dimension is the largest m for which $\Pi_H(m) = 2^m$. If we denote the vc dimension by d then sauer's lemma (Lemma 2.4 in the boosting book) states that

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$$

THe picture log of the growth function as a function of the number of examles.

4.5 VC dimension examples

In 2d.
Rectangles. Straight lines. Convex Polygons.
In higher dimensions
Planes. Radon's theorem.