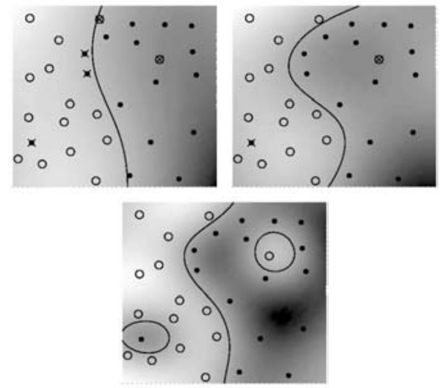
CSCE 990 Lecture 4: Statistical Learning Theory*

Stephen D. Scott

January 26, 2006

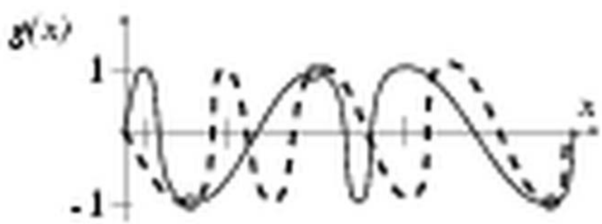
Introduction

- In Chapter 3, we discussed the need for restricting the class of functions $\mathcal{F}$ we choose from when minimizing $R_{\text{emp}}[f]$



Introduction
(cont'd)

- Put another way, simply minimzing $R_{\text{emp}}[f]$ doesn't necessarily minimize $R[f]$



- We will quantify the "expressiveness" or "richness" of $\mathcal{F}$ via its Vapnik-Chervonenkis (VC) dimension h, allowing us to bound $R[f]$ with probability at least $1-\delta$:

$$R[f] \leq R_{\text{emp}}[f] + \sqrt{\frac{1}{m}\left(h\left(\ln\frac{2m}{h}+1\right)+\ln\frac{4}{\delta}\right)} ,$$

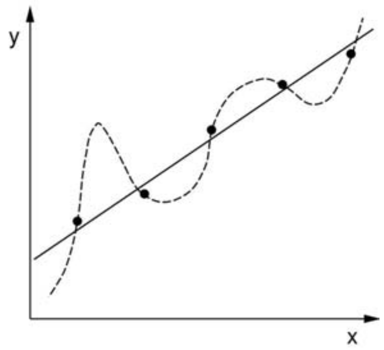
where m is training sample size

Outline

- Overfitting and the need for bias
- Expected risk minimization
- Law of large numbers
- Consistency and uniform convergence
- Vapnik-Chervonenkis dimension
- Aside: Structural risk minimization
- VC dimension of large-margin hyperplanes
- Example
- Sections 1.3, 5.1–5.4, 5.5.3–5.5.6, 5.6–5.7

An Example

- Consider two fits to m observations $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{X}, \mathcal{Y} = \mathbb{R}$:



- Restricting (biasing) our set of models to linear will always lead to simple explanations of S, though maybe not always good ones
- Allowing our set of models to be degree m polynomials will always lead to perfect explanations of S, but the models will have high variance in between data points
- This is the bias-variance dilemma, aka the issue of avoiding underfitting and overfitting

Empirical Risk Minimization

- Recall that our ultimate goal is to minimize the expected risk:

$$R[f] = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y, f(x)) \, d\mathrm{P}(x, y)$$

where commonly $\mathcal{Y} = \{+1, -1\}$ and $c(x, y, f(x)) = (1/2)|f(x) - y|$

- We don't know $\mathrm{P}(x, y)$, so we employ empirical risk minimization (ERM), and choose $f \in \mathcal{F}$ (where $\mathcal{F}$ is appropriately restricted) to minimize

$$R_{\mathrm{emp}}[f] = \frac{1}{m} \sum_{i=1}^{m} c(x_i, y_i, f(x_i))$$

Law of Large Numbers

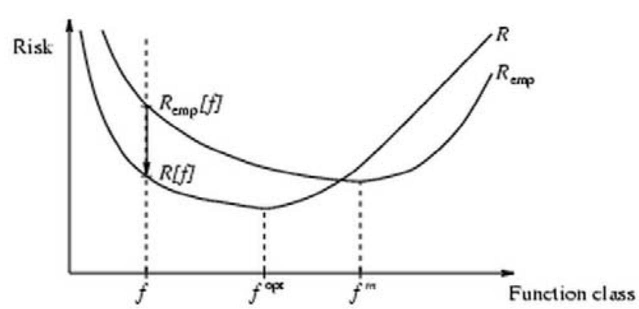
- For iid examples (x_i, y_i) and a fixed function f, the loss $\zeta_i = c(x_i, y_i, f(x_i))$ are also iid random variables
- In particular, if $c(x_i, y_i, f(x_i)) = (1/2)|f(x_i) - y_i|$, then $\zeta_i \in \{0, 1\}$ and are called Bernoulli trials
- Can apply Chernoff bound to show how quickly an empirical mean converges to its expectation:

$$\mathrm{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} \zeta_i - \mathbf{E}[\zeta] \right| \geq \epsilon \right) \leq 2 \exp\left(-2m\epsilon^2\right)$$

- I.e. for a fixed f, as the sample grows, the empirical risk converges exponentially fast to the true risk
- A more general form holds for $\zeta \in [a, b]$

Consistency

- Problem: f is not fixed!
  - Instead, we are choosing f to minimize $R_{\mathrm{emp}}[f]$
  - No longer have independent Bernoulli trials



- What we really want is more subtle: as $m \to \infty$, want $f^m = \mathrm{argmin}_{f \in \mathcal{F}} R_{\mathrm{emp}}[f]$ to also minimize $R[f]$
  - I.e. in the limit, f^m's training error matches its test error

Uniform Convergence

- In other words, want convergence of $R_{\text{emp}}[f]$ towards $R[f]$ to be uniform over all $f \in \mathcal{F}$
  - We'll come back to this later

- Let f^m be the function from $\mathcal{F}$ minimizing R_{emp} and let f^{opt} be the one minimizing R. Then $\forall\, f \in \mathcal{F}$

$$R[f] - R[f^{\text{opt}}] \geq 0 \ ,$$
$$R_{\text{emp}}[f] - R_{\text{emp}}[f^m] \geq 0$$

- Thus

$$R[f^m] - R[f^{\text{opt}}] \geq 0 \ , \qquad (1)$$
$$R_{\text{emp}}[f^{\text{opt}}] - R_{\text{emp}}[f^m] \geq 0 \qquad (2)$$

Uniform Convergence
(cont'd)

- Sum these and get

$$\begin{aligned} 0 &\leq R[f^m] - R_{\text{emp}}[f^m] + R_{\text{emp}}[f^{\text{opt}}] - R[f^{\text{opt}}] \\ &\leq \sup_{f \in \mathcal{F}} \left(R[f] - R_{\text{emp}}[f] \right) + R_{\text{emp}}[f^{\text{opt}}] - R[f^{\text{opt}}] \end{aligned}$$

- Since f^{opt} is a fixed function, we can apply our earlier results that say $R_{\text{emp}}[f^{\text{opt}}]$ approaches $R[f^{\text{opt}}]$ as $m \to \infty$

- Also, if we have uniform convergence (from above), then

$$\sup_{f \in \mathcal{F}} \left(R[f] - R_{\text{emp}}[f] \right) \xrightarrow{\text{P}} 0 \text{ as } m \to \infty$$

  (converges in probability; see p. 130)

- Thus in the limit, LHSs of (1) and (2) converge to 0, $R[f^m]$ is not larger than $R_{\text{emp}}[f^m]$, and ERM works

- UC also necessary for ERM (Theorem 5.3)

Vapnik-Chervonenkis Dimension

- Under what circumstances do we get uniform convergence?
  - I.e. what restrictions on $\mathcal{F}$ and m?

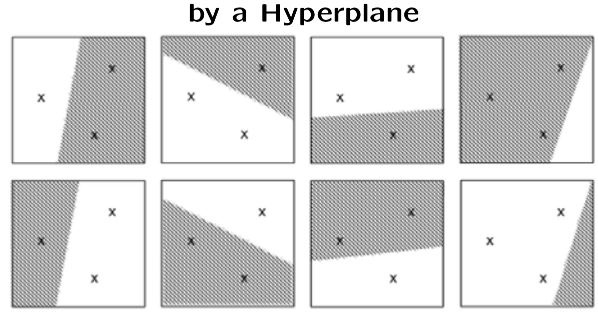
- There are many ways to quantify the "richness" of $\mathcal{F}$

- We will focus on the Vapnik-Chervonenkis (VC) dimension

Defn: A dichotomy of a set S is a partition of S into two disjoint subsets, i.e. into a set of $+$ patterns and a set of $-$ patterns

Defn: A set of instances S is shattered by set of functions $\mathcal{F}$ if and only if for every dichotomy of S there exists some function $f \in \mathcal{F}$ consistent with this dichotomy

Example: Three Instances Shattered by a Hyperplane



The Vapnik-Chervonenkis Dimension (cont'd)

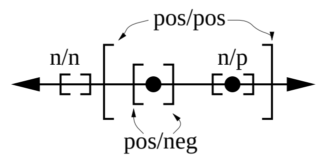
Defn: The Vapnik-Chervonenkis dimension h of $\mathcal{F}$ defined over $\mathcal{X}$ is the size of the largest finite subset of $\mathcal{X}$ shattered by $\mathcal{F}$. If arbitrarily large finite sets of $\mathcal{X}$ can be shattered by $\mathcal{F}$, then $h \equiv \infty$.

- So to show that $h = d$, must show there exists some subset $\mathcal{X}' \subset \mathcal{X}$ of size d that $\mathcal{F}$ can shatter and show that there exists no subset of $\mathcal{X}$ of size $> d$ that $\mathcal{F}$ can shatter

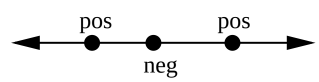
- Note that $h \leq \log_2 |\mathcal{F}|$ (why?)

VCD Example: Intervals on $\mathbb{R}$

- Let $\mathcal{F}$ be the set of closed intervals on the real line (each $f \in \mathcal{F}$ is a single interval), $\mathcal{X} = \mathbb{R}$, and a point $x \in \mathcal{X}$ is positive iff it lies in the interval defined by $f \in \mathcal{F}$



Can shatter 2 pts, so VCD >= 2



Can't shatter any 3 pts, so VCD < 3

- Thus $h = 2$

- In general, VCD of d-dimensional boxes is $2d$

VCD of Hyperplanes

(a) (b)

- Can't shatter (b), so what is lower bound on VCD?

- What about upper bound?

- In general, VCD of d-dimensional hyperplanes is $d + 1$

Putting it Together

- It turns out that if $\mathcal{F}$ has finite VCD then we can get uniform convergence and use ERM

- Skipping the proofs, one can show that for all $f \in \mathcal{F}$, with probability at least $1 - \delta$

$$R[f] \leq R_{\text{emp}}[f] + \sqrt{\frac{1}{m}\left(h\left(\ln\frac{2m}{h} + 1\right) + \ln\frac{4}{\delta}\right)} \tag{3}$$

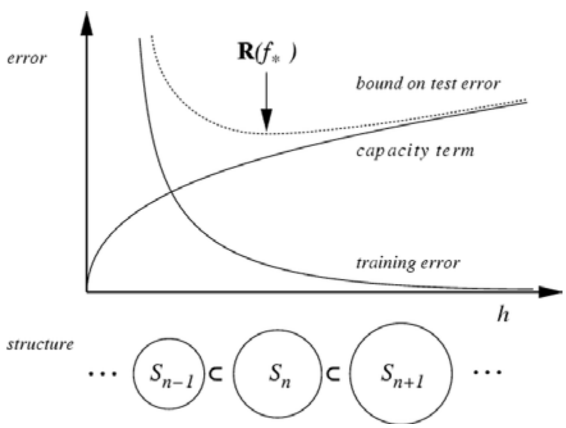
where m is the sample size

- Thus we have a tradeoff between low error on the training set and low VCD h

- Why the dependence on δ?

Aside: Structural Risk Minimization

- We can work with the tradeoff between R_{emp} and h via structural risk minimization (SRM)
- First decompose $\mathcal{F}$ into nested subsets of functions $S_1 \cdots S_{n-1} \subset S_n \subset S_{n+1} \cdots$ such that $h_1 < \cdots < h_{n-1} < h_n < h_{n+1} < \cdots$
- For each S_i, find the $f_i \in S_i$ minimzing R_{emp}
- Choose the f_i that minimizes (3)

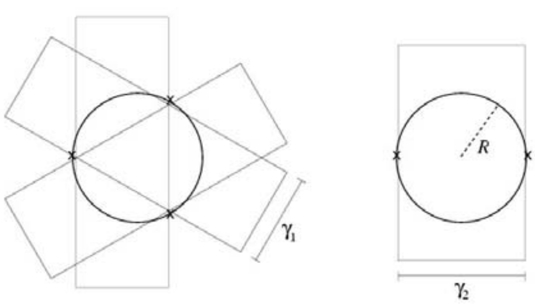


Back to SVMs

- What will the VCD be of our SVMs?
- Can we apply (3) to our results?

Back to SVMs
(cont'd)

- Recall that our SVMs not only find a hyperplane, but a large margin hyperplane



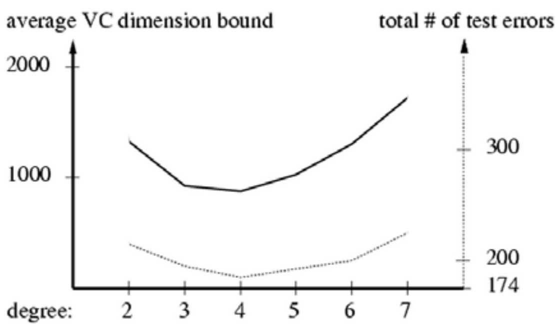
T5.5 Consider hyperplanes $\langle \mathbf{w}, \mathbf{x} \rangle = 0$ that are normalized such that $\min_{1 \le i \le r} |\langle \mathbf{w}, \mathbf{x}_i \rangle| = 1$ for some set of points $X^* = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ (i.e. the hyperplanes are in canonical form). Then the set of decision functions defined on X^* such that $\|\mathbf{w}\| \le \Lambda$ has VC dimension at most $R^2\Lambda^2$, where R is the radius of the smallest sphere centered at the origin and containing X^*.

Back to SVMs
(cont'd)

- Thus can substitute $R^2\Lambda^2$ for h in (3)
- Sort of (not exactly) motivates minimzing $\|\mathbf{w}\|$ in SVMs (see p. 142)
- Minimizing $\|\mathbf{w}\|$ corresponds to maximizing margin
  - This is our regularization term
- Can extend result to where ball is not centered at origin (by adding offset b) and to the entire input domain $\mathcal{X}$

Example

- Application of polynomial classifiers of degrees 2–7 to character recognition
- Data are separable for all degrees, so $R_{\text{emp}} = 0$ in all cases
- Ran 10 tests on different data sets, computed average VCD bound from T5.5 and average number of errors on independent test set



- VCD bound closely matches test error

Topic summary due in 1 week!