Definitions

- \( x \in \mathcal{X} \) (input or instance)
- \( d = c(x), c \in \mathcal{C} \) (label or concept) (consider binary labels)
- \( S = ((x(1),d(1), \ldots,x(m),d(m)) \) (sample drawn iid from some unknown distribution \( D \))
- \( h = \mathcal{L}(S,C) \) takes a sample \( S \) and chooses a hypothesis concept class consistent with \( S \) (i.e. \( h(x(i)) = d(i) \))
- \( P(c \neq h) = D(c \Delta h) \)
PAC Learning

Probably Approximately Correctly: Let $C$ be a concept class over $\mathcal{X}$. $C$ is PAC learnable if there exists $\mathcal{L}$ such that for $c \in C$, for every distribution $D$ on $\mathcal{X}$, for all $0 < \varepsilon < 1/2$, $0 < \delta < 1/2$, then with probability of at least $1 - \delta$, $\mathcal{L}$ outputs hypothesis $h \in C$ such that $P(c \neq h) < \varepsilon$.

If $\mathcal{L}$ runs in time polynomial in $1/\varepsilon$, $1/\delta$ and $m$ we say $C$ is efficiently PAC learnable. $\varepsilon$ is error parameter and $\delta$ is confidence parameter.
Learning Example

- Concept class: (Intervals $[0,a]$ represents positive points)
- Inputs drawn from positive examples only: hypothesis concept $h = [0, \max(x(1), \ldots, x(m))]$
- In PAC learning setting we have
  \[ P((c \Delta h) > \varepsilon) = (1 - \varepsilon)^m < \exp(-\varepsilon m) < \delta \text{ or } m > \frac{1}{\varepsilon} \log(1/\delta) \]
- Second example: Axis aligned rectangles
- For general setting given $\varepsilon$, error parameter $\delta$, confidence parameter, and $h$ VC dimension of concept class, and $m$, number of training examples find bounds relating on error rates relating these quantities
VC dimension

- Consider function classes where each function labels each input as 1 or –1.
- A set of $m$ points is shattered by function class if the function class represents all $2^m$ possible labelings of the points.
- The VC dimension of a function class is the largest cardinality of points that is shattered by the function class. Example: linear threshold functions in Euclidean $n$ space has VC dimension of $n+1$.
- The VC dimension measures the complexity of the function class.
Growth functions and numbers

- **Growth function**: $\Pi_C(S) = \{c \cap S : c \in C\}$, note that $\Pi_C(S) \subseteq \{0,1\}^S$ (power set). If equality, then $C$ shatters $S$.

- **Growth number**: $\Pi_C(m) = \max_{|S|=m} |\Pi_C(S)|$

- **VC dimension** $\text{VC}(C) = \max m$ such that $\Pi_C(m) = 2^m$

If no number exists, then VC dimension is infinite.
VC dimension examples

- Homogenous Linear Threshold Functions: n
- Linear Threshold Functions: n+1
- Quadratic Threshold Functions: \((n+1)(n+2)/2\)
- One closed interval: 2
- Closed intervals: \(\infty\)
- Axis aligned rectangles: 2n
Learning Theorem

- Lemma 1: If $\text{VC}(C) = d$ then for any $m$ we have that $\Pi_C(m) \leq C(m+1,d+1)/2$
- Lemma 2: If $\text{VC}(C) = d$ and $m > d$, then $\Pi_C(m) \leq (em/d)^d$.
- Theorem: Let $C$ be any concept class with $\text{VC}(C) = d$. Let $\mathcal{L}$ be any algorithm that takes as input a set $S$ of $m$ labeled examples drawn from an unknown arbitrary distribution $D$ and produces $h \in C$ that is consistent with labeled examples. Then $\mathcal{L}$ is an efficient PAC learning algorithm where $m \geq k_1((1/ \varepsilon) \log(1/ \delta) + (d/ \varepsilon) \log(1/ \varepsilon))$. 
Learning Theorem Comments

- Sketch of proof: \( P(h \text{ consistent, error}(h) > \varepsilon) < \delta \). LHS depends on two terms growth function which grows polynomially as \( m \) and a second term that decays exponentially as \( m \) (Hoeffding bound)

\[
P(h \text{ consistent, error}(h) > \varepsilon) \leq \left(\frac{2em}{d}\right)^d 2\left(2^{-\varepsilon m/2}\right) < \delta
\]

then solve for \( m \).

- Lower bound: need at least \( k_2 \frac{d}{\varepsilon} \) examples.

- Results similar to LLN and typical sequence results in information theory

- Results are distribution free and bounds are not tight
Learning Theory Applications

• Learning finite concept class
• Learning boolean functions
• Learning capacity of feedforward neural networks
• Learning capacity of kernel machines
• Extensions to analog outputs
Vapnik proved an important result giving an upper bound on generalization error. Let $d$ be the VC dimension of a function class and $m$ the number of training examples, then

$$J(w) \leq J_{\text{emp}}(w) + \sqrt{\frac{d(\ln(2m/d)+1) - \ln(\eta/4)}{m}}$$

holds with probability $1-\eta$, where the second term is a confidence term.

Bound independent of distribution and based on drawing iid samples.

Lower bounds can also be established.

Want to establish consistency of $J_{\text{emp}}(w)$ to converge to $J(w)$. 

**Relationship between empirical error and generalization error**

Vapnik proved an important result giving an upper bound on generalization error. Let $d$ be the VC dimension of a function class and $m$ the number of training examples, then

$$J(w) \leq J_{\text{emp}}(w) + \sqrt{\frac{d(\ln(2m/d)+1) - \ln(\eta/4))}{m}}$$

holds with probability $1-\eta$, where the second term is a confidence term.

Bound independent of distribution and based on drawing iid samples.

Lower bounds can also be established.

Want to establish consistency of $J_{\text{emp}}(w)$ to converge to $J(w)$. 

Consider a set of growing function classes of increasing complexity

\[ C_1 \subset C_2 \ldots C_k \subset C_{k+1} \subset \ldots \]

- Training error
- Bound on test error
- Confidence term
- Error
- VC dimension
SRM for SVM

For support vector machines we can consider concept classes of the following form:

\[ C_i = \{ w^T \phi(x) + b : \|w\| \leq c_i \} \text{ where } c_1 < c_2 < \ldots < c_i < \ldots \]

Vapnik showed that hyperplanes satisfying \( \|w\| \leq c \) have VC dimension \( h \) that is upper bounded by

\[ h \leq \min \left( \lfloor r^2 c^2 \rfloor, d \right) + 1 \]

Where \( r \) is the radius of the largest ball surrounding all inputs, \( d \) is the dimensionality of feature space, and \( \lfloor x \rfloor \) is the integer part of \( x \).
SVM and Learning Theory comments

- For SVMs like feedforward networks there are several parameters to preset.
  - Cost term: C
  - Type of kernel and kernel parameters
- To perform SRM can use
  - Trial and error
  - Crossvalidation
  - Bayesian learning methods
- Occam’s razor: The simplest explanation is the best.
More General Approaches to Learning

- Concentration Inequalities: Hoeffding’s Inequality, McDiarmid
- Uniform Convergence and Capacity: Rademacher complexity
- Learning for kernel based classes
Relationship between empirical error and generalization error
Vapnik proved an important result giving an upper bound on generalization error.

\[ J(w) \leq J_{\text{emp}}(w) + J_{\text{model}}(w) \]

- VC dimension theory
- Margin and slack variable bounds
- Rademacher complexity
- Extensions to regression and novelty detection
- Linking statistical learning and Bayesian learning
More General Approaches to Learning

- Concentration Inequalities: Hoeffding’s Inequality, McDiarmid
- Uniform Convergence and Capacity: Rademacher complexity
- Learning for kernel based classes
Concentration Inequalities

Random vectors are often concentrated around their means and a goal of concentration inequalities is to bound these probabilities.

- **Markov’s Inequality:** For any nonnegative random variable (RV) $X$, $P(X \geq \varepsilon) \leq E(X)/\varepsilon$.
- If $X$ is a zero mean RV with $X \in [a,b]$ then for $s \geq 0$, $E(\exp(sX)) \leq \exp(s^2(b-a)^2/8)$
- **Hoeffding’s Inequality:** Let $X_i$ be a sequence of iid RVs that are bounded $f(X_i) \in [a,b]$ and $S_m = \sum_{i=1}^m f(X_i)$, then $P(|S_m - E(S_m)| \geq \varepsilon) \leq \exp(-2\varepsilon^2/(m(b-a)^2))$
McDiarmid’s Inequality: Let $X_i$ be a sequence of ind RVs that take values in a set $A$ and assume that $f: A^m \rightarrow \mathbb{R}$ that satisfies

$$\sup_{x_1, \ldots, x_m, z_i \in A} |f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, z_i, \ldots, x_m)| \leq c_i$$

then for all $\varepsilon > 0$

$$P(|f(X_1, \ldots, X_m) - E(f(X_1, \ldots, X_m))| \geq \varepsilon) \leq \exp(-\frac{2\varepsilon^2}{\sum_{i=1}^{m} (c_i)^2})$$
Rademacher Theory

- **Rademacher complexity**: Let $\mathcal{F}$ be a class of real valued functions on $A$, $X_i$ be a sequence of iid RVs and $\sigma(i)$ be an iid sequence of Rademacher RVs (Bernoulli ($-1, 1$) RVs with $p = 0.5$). Define the Rademacher average as
  
  \[ R_m(\mathcal{F}) = E(\sup_{f \in \mathcal{F}} 1/m \sum_{i=1}^{m} \sigma(i) f(X_i)) \]

- **Empirical Rademacher complexity**
  
  \[ \hat{R}_m(\mathcal{F}) = E(\sup_{f \in \mathcal{F}} 1/m \sum_{i=1}^{n} \sigma(i) f(X_i)|X_1, \ldots, X_m) \]

- **Rademacher complexity** measures complexity of class of functions and is related to VC dimension
Estimating mean of a Random Vector

Given a random sample \( S = \{x_1...,x_m\} \) can we accurately bound the sample mean \( \phi_S = 1/m \sum_{i=1:m} \phi(x_i) \) by the ensemble mean

\[
E(\phi(X)) = \int \phi(x) dP(x) .
\]

Define \( g(S) = \|\phi_S - E(\phi(X))\| \)

Apply McDiarmid’s inequality to RV \( g(S) \) by bounding this inequality when \( x_i \) is replaced by \( z_i \) to give \( S \)

\[
|g(S) - g(S)| \leq \|\phi_S - E(\phi(X))\| - \|\phi_S - E(\phi(X))\| \leq \| \phi_S - \phi_S \|
\]

\[
= 1/m \| \phi(x_i) - \phi(z_i) \| \leq 2C/m
\]

where \( C = \sup x \in A \| \phi(x) \| \) obtaining

1. \( P( g(S) - E(g(S)) \geq \varepsilon) \leq \exp \left( -2m\varepsilon^2/4R^2 \right) \)
Estimating mean continued

2. $E[g(S)] = E[||\phi_S - E(\phi(X))||] = E_S[||\phi_S - E_S(\phi_S)||]$ (S ind. of S)
   
   $= E_S[|| E_S[\phi_S - \phi_S]|| \leq E_S[||\phi_S - \phi_S||]$ (triangle ineq.)
   
   $= E_{\sigma S_S}[|| 1/m \sum_{i=1:m} \sigma(i)\phi(x_i)-1/m \sum_{i=1:m} \sigma(i)\phi(z_i)||]$
   $\leq 2E_{\sigma S} [|| 1/m \sum_{i=1:m} \sigma(i)\phi(x_i)||]$ (triangle ineq.)
   
   $= 2/m E_{\sigma S} [ \langle \sum_{i=1:m} \sigma(i)\phi(x_i), \sum_{j=1:m} \sigma(j)\phi(x_j) \rangle]^{1/2}$
   $\leq 2/m [E_{\sigma S} (\sum_{i,j=1:m} \sigma(i) \sigma(j) K(x_i, x_j))]^{1/2}$ (Jensen’s ineq.)
   
   $= 2/m [E_S (\sum_{i,j=1:m} K(x_i, x_j))]^{1/2} \leq 2C/m^{1/2}$

Two key steps: augmenting $S$ by $S$ which is independent of $S$ with $|S| = |S|$ and introducing $\sigma(i)$ which are Rademacher RVs
Estimating mean continued

- Set $\exp(-2m \varepsilon^2/(4R^2)) = \delta$ and solve for $\varepsilon$.
- Combine equations 1. and 2. together to get that
  \[ P(g(S) \leq C/m^{1/2} \left(2 + (2\ln(1/\delta))^{1/2}\right)) \geq 1 - \delta \]
- Note that bound is independent of dimension of feature space. The bound involves using concentration inequalities, augmentation of samples, and Rademacher RVs.
Rademacher Properties

Let $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_m$, and $\mathcal{G}$ be classes of real functions. Let $\text{conv}(\mathcal{F})$ denote convex combinations of elements of $\mathcal{F}$.

- If $\mathcal{F} \subseteq \mathcal{G}$, then $R_m(\mathcal{F}) \leq R_m(\mathcal{G})$
- $R_m(\mathcal{F}) = R_m(\text{conv}(\mathcal{F}))$
- For every $c \in \mathbb{R}$, $R_m(c\mathcal{F}) = |c| R_m(\mathcal{F})$
- If $\phi: \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant $L_\phi$ and satisfies $\phi(0)$ then $R_m(\phi \circ \mathcal{F}) \leq 2 L_\phi R_m(\mathcal{F})$
- For any uniformly bounded function $h$, $R_m(\mathcal{F}+h) \leq R_m(\mathcal{F}) + ||h||_\infty / (n)^{1/2}$
- For $1 \leq q < \infty$, let $\mathcal{L}_{\mathcal{F},h,q} = \{|f-h|^q \mid f \in \mathcal{F}\}$, where $h$ is uniformly bounded. If $||f-h||_\infty \leq 1$ for every $f \in \mathcal{F}$, then $R_m(\mathcal{L}_{\mathcal{F},h,q}) \leq 2q \left( R_m(\mathcal{F}) + ||h||_\infty / (n)^{1/2} \right)$
- $R_m(\sum_{i=1:k} \mathcal{F}_i) \leq \sum_{i=1:k} R_m(\mathcal{F}_i)$
Computing Rademacher Complexity

- Rademacher complexity related to VC dimension, Gaussian complexity, and maximum discrepancy.
- Rademacher bounds (statistical complexity) can be substantially better than VC dimension bounds (combinatorial complexity).
- Rademacher complexity is often difficult to determine, but can estimate complexity from empirical Rademacher complexity and use results from previous slide to consider complexity of different class of functions.
Pattern Classification Bounds

Let $P$ be a probability distribution on $X \times D$ with $D = \{\pm 1\}$, let $F$ be a set of $\{\pm 1\}$ functions defined on $X$, and let $S$ be a sample of $m$ training examples drawn according to $P^m$. Then with probability at least $1 - \delta$ every function $f \in F$ satisfies

$$P(D \neq f(X)) \leq P_m(D \neq f(X)) + R_m(F)/2 + (\ln(1/\delta))^{1/2}/(2m))$$

where $P_m(D \neq f(X)) = (1/m) \sum_{i=1:m} 1_{S}(D_i \neq f(X_i))$.

Let $L(D, f(X)) = 1(D \neq f(X))$, then proof involves bounding $\sup_{h \in L \land F} (E(h) - E_m(h))$ where $E_m(h)$ denotes empirical average and then using McDiarmid’s inequality and Rademacher complexity to get the second and third terms on the RHS of inequality.
Kernel Application

Fix $\gamma > 0$, $\delta \in (0,1)$, let $\mathcal{F}$ be a class of functions from $X \times D$ to $\mathbb{R}$ with $D = \{\pm 1\}$ given by $f(x,d) = -dg(x)$ where $g$ is a linear function in a kernel-defined feature space with norm at most 1. Let $S$ be a sample of training examples drawn from $P$. Then with probability at least $1 - \delta$ over samples of size $m$ we have that

$$P(D \neq \text{sgn}(g(X))) \leq \frac{1}{m\gamma} \sum_{i=1:m} \xi_i + \frac{4}{m\gamma} (\text{tr}(K))^{1/2} + 3\left(\ln\left(\frac{2}{\delta}\right)\right)^{1/2} / (2m)$$

where $K$ is kernel matrix from training set and $\xi_i = \xi((x_i,d_i),\gamma,g)$
Learning Theory References