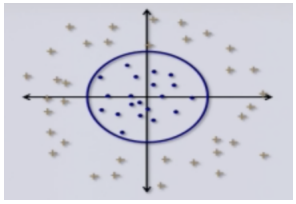
ECE 2372 - Pattern Recognition

Supplementary reading for today's lecture: "Learning from Data" Chapter 2.1.1 and 2.1.2

So far, we talked about:

- Linear Discriminant Analysis
- Logistic Regression
- Perceptron Learning Algorithm
- Maximum margin hyperplanes

Pictured data set is not linearly separable but can be in a higher dimension with a transform:



$$\Phi(x) = \begin{bmatrix} 1 \\ x(1) \\ x(2) \\ x(1)x(2) \\ x(1)^2 \\ x(2)^2 \end{bmatrix}$$

This dataset is linearly separable after applying such transformation with w=$[-1, 0, 0, 1, 1]^T$

Fundamental Tradeoff: By mapping the data to a higher-dimensional space, the set of linear classifiers becomes a **"richer set"**.

$$\text{Richer set of hypothesis} \implies \begin{cases} \hat{R}_n(h^*) & \downarrow \\ \hat{R}_n(h^*) - R(h^*) & \uparrow \end{cases}$$

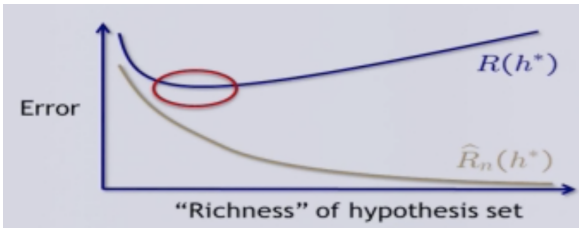


Figure 1: Tradeoff

Measure for "richness":

When can we have confidence that $\hat{R}_n(h^*) \approx R(h^*)$ where h^* is chosen from an **infinite set** $\mathcal{H}$.

- For a single hypothesis,

$$\mathbb{P}[\ \mid \hat{R}_n(h) - R(h) \mid\ > \epsilon\ ] \leq 2e^{-2\epsilon^2 n}$$

- For $m =\mid \mathcal{H} \mid$ hypothesis, and $h^* \in \mathcal{H}$

$$\mathbb{P}[\ \mid \hat{R}_n(h^*) - R(h^*) \mid\ > \epsilon\ ] \leq 2me^{-2\epsilon^2 n}$$

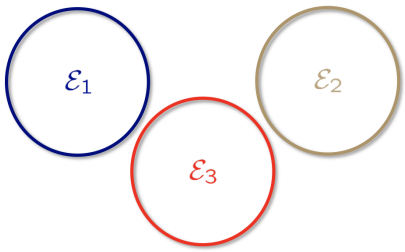
Where did m come from? Union bound:

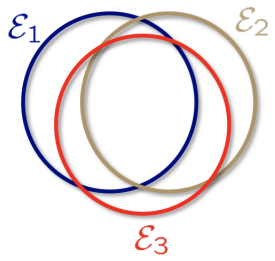
$$\mathbb{P}[\varepsilon_1 \cup ... \cup \varepsilon_m] \leq \mathbb{P}[\varepsilon_1] + ... + \mathbb{P}[\varepsilon_m]$$

Here the events we are bounding:

$$\varepsilon_j = \mid \hat{R}_n(h_j) - R(h_j) \mid\ > \epsilon$$

So pictorially, possibilities for these bad events:





One thing clear from this picture is we can improve on m is there is an overlap between "bad events". In other words get a better bound than union suggests. It turns out in reality, we are much closer to the situation on right figure, there is tremendous overlap between bad events.

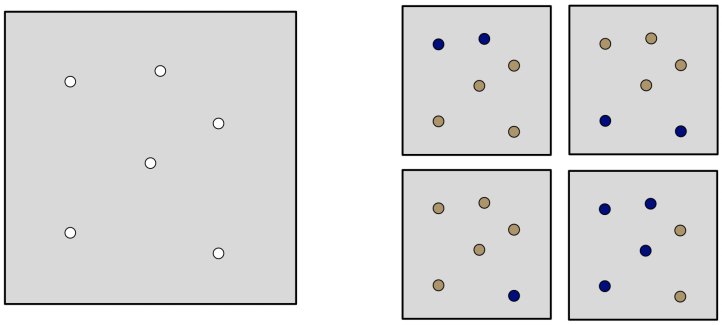
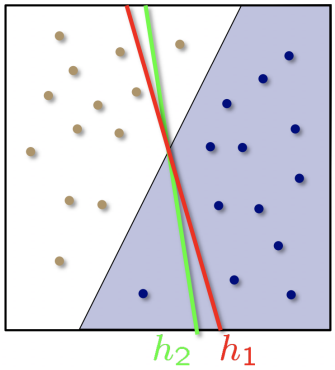


Figure 2: Dichotomies



$$R(h_1) \approx R(h_2) \quad \text{and} \quad \hat{R}_n(h_1) \approx \hat{R}_n(h_2)$$

$$\implies | \hat{R}_n(h_1) - R(h_1) | \approx | \hat{R}_n(h_2) - R(h_2) |$$

What can we substitute m with? These events are very overlapping, using the union bound is not the best idea.

- Small changes into hypothesis may lead into small changes in true risk
- Rather than considering all possible hypothesis we have in $\mathcal{H}$, we will consider a finite set of input points $x_1, ..., x_n$ and "combine" hypothesis that result in the same labeling.
  - we call a particular labeling of $x_1, ..., x_n$ a **dichotomy**

Hypotheses vs dichotomies:

Hypotheses

- $h : \mathcal{X} \to \{-1, +1\}$
- Number of hypothesis is $| \mathcal{H} |$ potentially infinite
- $| \mathcal{H} |$ (or m) is a poor way to measure "richness" of $\mathcal{H}$.

Dichotomies

- $h : \{x_1, ..., x_n\} \to \{-1, +1\}$
- Number of dichotomies $\mid \mathcal{H}(x_1, ..., x_n) \mid$ is at most 2^n (unique labellings).
- This is a good candidate for replacing $\mid \mathcal{H} \mid$ as a measure of "richness".

The growth function: A dichotomy is defined in terms of a particular $x_1, ..., x_n$.

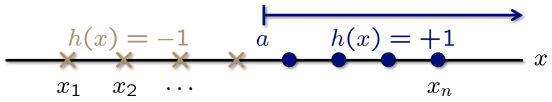
The growth function of $\mathcal{H}$ is defined as : $m_{\mathcal{H}}(n) = max_{x_1,...,x_n \in \mathcal{X}} \mid \mathcal{H}(x_1, ..., x_n) \mid$

$m_{\mathcal{H}}(n)$ counts the **most** dichotomies that can possibly be generated on n points.

One can show that $m_{\mathcal{H}}(n) \leq 2^n$, but it can potentially be much smaller.

Example 1: Positive rays

Candidate functions: $h : \mathbb{R} \to \{-1, +1\}$ such that $h(x) = sign(x - a)$ for some $a \in \mathbb{R}$.

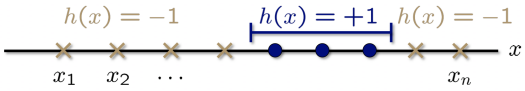


$$m_{\mathcal{H}}(n) = n + 1$$

Example 2: Positive intervals

Candidate functions: $h : \mathbb{R} \to \{-1, +1\}$ such that

$$h(x) = \begin{cases} +1 & \text{for} \quad x \in [a, b] \\ -1 & \text{otherwise} \end{cases}$$

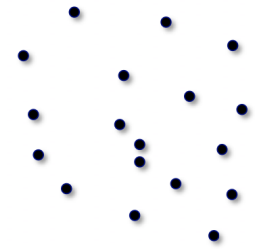


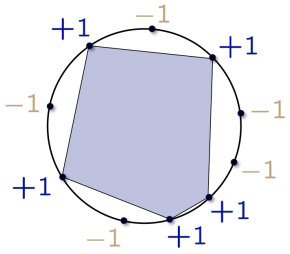
$$\begin{aligned} m_{\mathcal{H}}(n) &= \binom{n+1}{2} + 1 \\ &= \frac{1}{2}n^2 + \frac{1}{2}n + 1 \end{aligned}$$

Example 3: Convex sets

Candidate functions: $h : \mathbb{R}^2 \to \{-1, +1\}$ such that

$$\{x : h(x) = +1\} \text{ is convex}$$





Is there any labeling that you can't draw a convex shape around?

$$m_{\mathcal{H}}(n) = 2^n$$

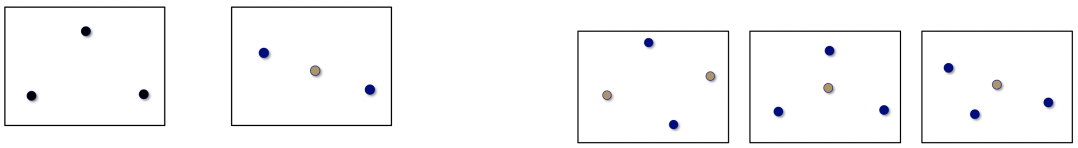
If $\mathcal{H}$ can generate all possible dichotomies on $x_1, ..., x_n$, then it is refereed as that $\mathcal{H}$ **shatters** $x_1, ..., x_n$.

Example 4: Linear classifiers Candidate functions: $h : \mathbb{R}^2 \to \{-1, +1\}$ such that

$$[h]\{x : h(x) = sign(\mathrm{w}^T x + b)\}$$

for some $\mathrm{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$.

- $m_{\mathcal{H}}(3) = 2^3$
- $m_{\mathcal{H}}(4) = 14$



Recap:

- Positive rays:$m_{\mathcal{H}}(n) = n + 1$
- Positive intervals: $m_{\mathcal{H}}(n) = \frac{1}{2}n^2 + \frac{1}{2}n + 1$
- Convex sets: $m_{\mathcal{H}}(n) = 2^n$
- Linear classifiers in $\mathbb{R}^2$:

$$\begin{aligned} m_{\mathcal{H}}(1) &= 2 \\ m_{\mathcal{H}}(2) &= 4 \\ m_{\mathcal{H}}(3) &= 8 \\ m_{\mathcal{H}}(4) &= 14 \\ m_{\mathcal{H}}(n) &= \quad ? \end{aligned}$$

Recap:

- Challenge: Number of hypothesis is $\mid \mathcal{H} \mid$ potentially infinite
- Better: Narrow the scope to the finite training set in order to replace easily infinite m. Dichotomies allow us that.
- $h : \{x_1, ..., x_n\} \mapsto \{-1, 1\} \implies 2^n$ different way of labeling, max! so dichotomy is the way of labeling THAT particular data set
- Hence, in general, $\mid \mathcal{H} \mid > \mid \mathcal{H}(x_1, ..., x_n) \mid$. In English number of hypothesis ¡ number of dichotomies
- So maybe a dichotomies are a better measure of "richness" of the set.
- And then we introduced the idea of "growth function that gets rid of the dependence of dichotomy to a particulars of our training set $x_1, ..x_n$.
- Growth function: $m_{\mathcal{H}}(n) = max_{x1,...,x_n \in \mathcal{X}} \mid \mathcal{H}(x_1, ..., x_n) \mid$.

Recall

$$\mathbb{P}[\mid \hat{R}_n(h^*) - R(h^*) \mid > \epsilon] \leq 2me^{-2\epsilon^2 n}$$

Another way to write this, if you pick a δ then we can guarantee that with probability at least $1 - \delta$

$$R(h^*) \leq \hat{R}_n(h^*) + \sqrt{\frac{1}{2n}\ log\ \frac{2m}{\delta}}$$

by setting $2me^{-2\epsilon^2 n} = \delta$ and solve for ϵ. If $m \propto e^n$, we have a problem...

No matter how big n gets $\sqrt{\frac{1}{2n}\ log\ \frac{2m}{\delta}}$ will never be smaller...

What if we replace with m with $m_{\mathcal{H}}(n)$? Suppose that for any $\delta \in (0,1)$, we can guarantee at least $1-\delta$

$$R(h^*) \leq \hat{R}_n(h^*) + \sqrt{\frac{1}{2n}\ log\ \frac{2m_{\mathcal{H}}(n)}{\delta}}$$

- If $m_{\mathcal{H}}(n) = 2^n$ then $\sqrt{\frac{1}{2n}\ log\ \frac{2m_{\mathcal{H}}(n)}{\delta}}$ is a constant
- If $m_{\mathcal{H}}(n)$ is a polynomial in n, $\sqrt{\frac{1}{2n}\ log\ \frac{2m_{\mathcal{H}}(n)}{\delta}}$ decays like $\sqrt{\frac{log\ n}{n}}$.

When is learning feasible?

Assuming that we are indeed allowed to substitute $m_{\mathcal{H}}(n)$ for m, we can argue that for a given set of hypothesis $\mathcal{H}$ learning is possible provided that $m_{\mathcal{H}}(n)$ is a polynomial. How do we know it is a polynomial?

Key idea: Break points

def'n: If no data set of size k can be shattered by $\mathcal{H}$, then k is a **break point** for $\mathcal{H}$.

$$m_{\mathcal{H}(k)} < 2^k$$

This also implies that if k is a break point, then so is any $k' > k$.

Examples of Break points

- Positive rays:$m_{\mathcal{H}}(n) = n+1$
  - break point: $k = 2$
- Positive intervals: $m_{\mathcal{H}}(n) = \frac{1}{2}n^2 + \frac{1}{2}n + 1$
  - break point: $k = 3$
- Convex sets: $m_{\mathcal{H}}(n) = 2^n$
  - break point: $k = \infty$
- Linear classifiers in $\mathbb{R}^2$:
  - break point: $k = 4$

If there exists any break point, then $m_{\mathcal{H}}(n)$ is polynomial in n

If no break points, then $m_{\mathcal{H}}(n) = 2^n$

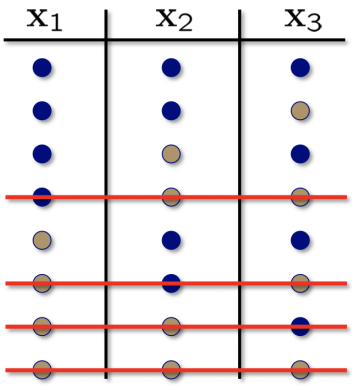
As soon as we have a single break point, this starts eliminating tons of dichotomies.

- We can show that $m_{\mathcal{H}}(n)$ is polynomial in n.
- We can show that $m_{\mathcal{H}}(n) \leq$ **some** polynomial
- Main approach will center around:
  - $B(n, k) :=$ maximum number of dichotomies on n points such that no subset of size k can be shattered by these dichotomies
  - Notice that this is a purely combinatorial quantity
  - By definition, $m_{\mathcal{H}}(n) \leq B(n, k)$

Example: how many dichotomies?

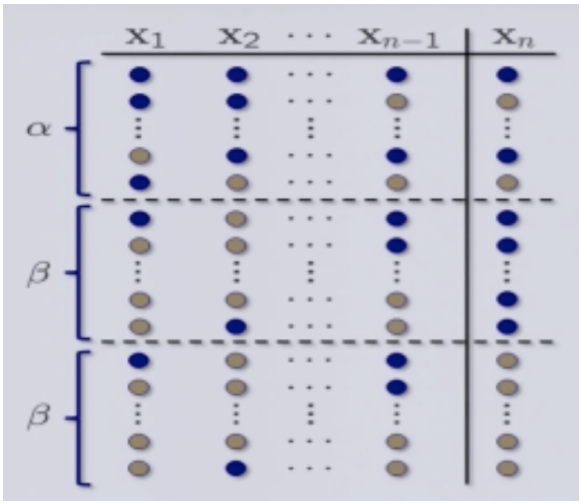
You are given a hypothesis set which has a break point of 2.

How many dichotomies can you get on 3 data points?



Summary: $B(n, k)$ is the combinatorial quantity that's an upper bound on the growth function for any possible set of classifiers.

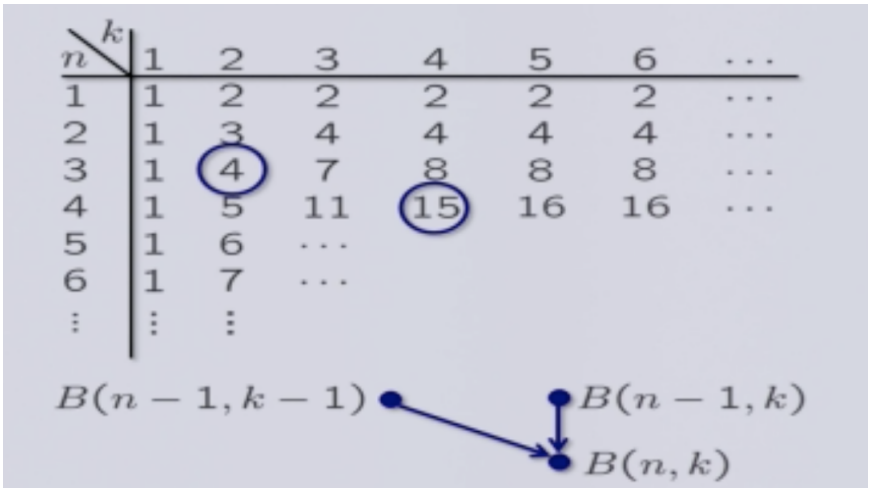
You can bound $B(n, k)$ recursively which is an algorithmic proof. We will just skip the analytical proof as it is pages and pages math. There is also a "proof by picture" for this which I like, you may review that from "Learning from Data" if you are interested...



$$\alpha + \beta \leq B(n-1,k)$$
$$\beta \leq B(n-1,k-1)$$

Hence,

$$B(n,k) \leq B(n-1,k) + B(n-1,k-1)$$



Analytical solution: $B(n,k) \leq \sum_{i=0}^{k-1} \binom{n}{i}$ You can prove that it is actually equal,

$$B(n,k) = B(n-1,k) + B(n-1,k-1)$$

but all we really need is an upper bound, so that is all we will prove here.

Proof by induction:

$$B(n,k) \leq B(n-1,k) + B(n-1,k-1)$$

- Base case

$$B(n,1) = 1$$

$$B(1,k) = \begin{cases} 1 & \text{if} \quad k = 1 \\ 2 & \text{otherwise} \end{cases}$$

- Inductive step
  - suppose the inequality is true for $B(n-1,k)$ and $B(n-1,k-1)$

$$\begin{aligned} B(n,k) &\leq \sum_{i=0}^{k-1} \binom{n-1}{i} + \sum_{i=0}^{k-2} \binom{n-1}{i} \\ &= 1 + \sum_{i=0}^{k-1} \binom{n-1}{i} + \sum_{i=1}^{k-1} \binom{n-1}{i-1} \\ &= 1 + \sum_{i=1}^{k-1} (\binom{n-1}{i} + \binom{n-1}{i-1}) \\ &= 1 + \sum_{i=1}^{k-1} \binom{n}{i} = \sum_{i=0}^{k-1} \binom{n}{i} \end{aligned}$$