ECE 0402 - Pattern Recognition

Lecture 7 on 2/7/2022

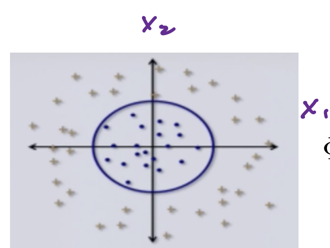
Supplementary reading for today's lecture: "Learning from Data" Chapter 2.1.1 and 2.1.2

So far, we talked about:

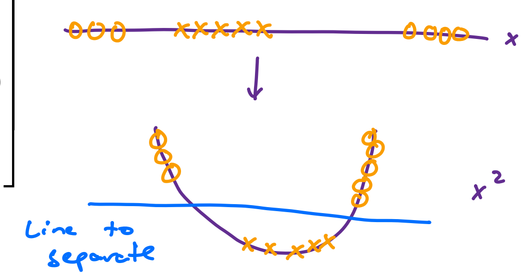
- Linear Discriminant Analysis
- Logistic Regression
- Perceptron Learning Algorithm
- Maximum margin hyperplanes

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

Best way to separate is with circle



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



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"**.

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

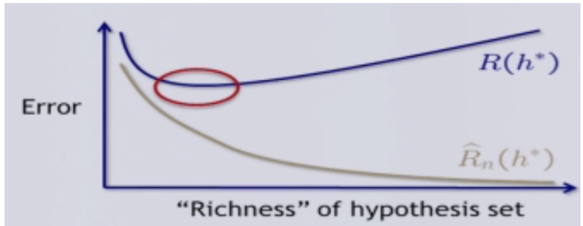


Figure 1: Tradeoff

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,

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

ERM

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

$$\underbrace{\mathbb{P}[|\ \hat{R}_n(h^*) - R(h^*)\ | > \epsilon]}_{\text{Bad event}} \ \leq \ 2me^{-2epsilon^2 n}$$

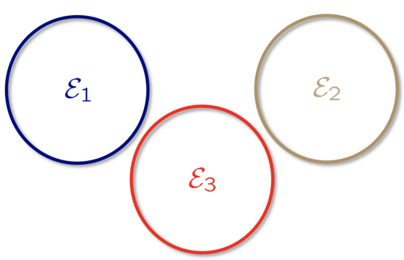
Where did m come from? Union bound:

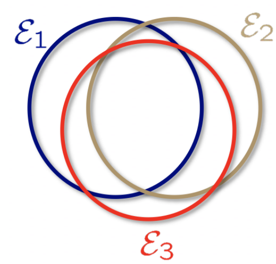
$$\mathbb{P}[\varepsilon_1 \cup ... \cup \varepsilon_m] \leq \mathbb{P}[\varepsilon_1] + ... + \mathbb{P}[\varepsilon_m]$$

Here the events we are bounding:

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

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.

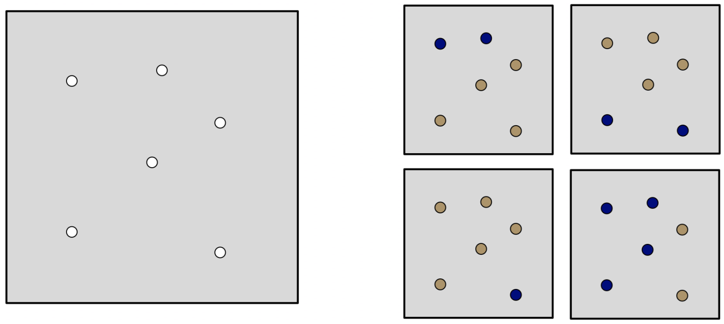
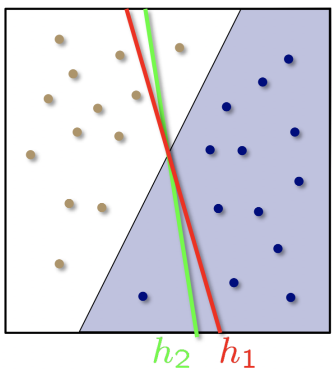


Figure 2: Dichotomies



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

$$\implies \overbrace{\mid \hat{R}_n(h_1) - R(h_1) \mid}^{\varepsilon_1} \approx \overbrace{\mid \hat{R}_n(h_2) - R(h_2) \mid}^{\varepsilon_2}$$

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**

How many ways can you label n data points (binary classification)?
→ 2^n (dichotomy)

Hypotheses vs dichotomies:

Hypotheses

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

Dichotomies

- $h : \{x_1, ..., x_n\} \rightarrow \{-1, +1\}$
- Number of dichotomies $| \mathcal{H}(x_1, ..., x_n) |$ is at most 2^n.
- Good candidate for replacing $| \mathcal{H} |$ 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 : $m_{\mathcal{H}}(n) = max_{x_1,...,x_n \in \mathcal{X}} \, | \mathcal{H}(x_1, ..., x_n) |$

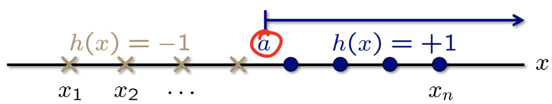
$m_{\mathcal{H}}(n)$ counts the **most** dichotomies that can possibly be generated on n points.

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

Example 1: Positive rays

$sign(x-a)$

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



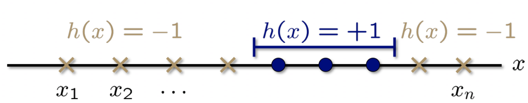
$m_H(n) \leq 2^n$

$$m_{\mathcal{H}}(n) = n + 1 \quad << 2^n$$

Example 2: Positive intervals

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

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

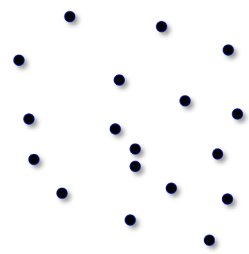


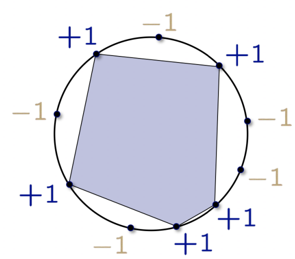
$$m_{\mathcal{H}}(n) = \binom{n+1}{2} + 1$$
$$= \frac{1}{2}n^2 + \frac{1}{2}n + 1$$

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 $b \in \mathbb{R}$.

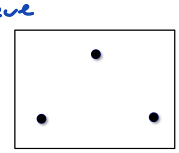
Impression of 2^4

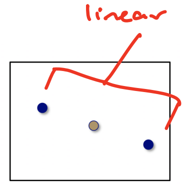
BUT can't be labeled

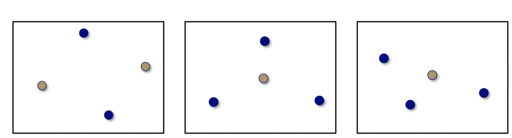
- $m_{\mathcal{H}}(3) = 2^3$
- $m_{\mathcal{H}}(4) = 14$

Can't classify with linear classifier

Can achieve all 8 labelings for this set







Recap:

$2^n - 2^{n-3}$

- 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 $\mathbb{R}^2$:

$$\begin{aligned}
m_{\mathcal{H}}(1) &= 2 \\
m_{\mathcal{H}}(2) &= 4 \\
m_{\mathcal{H}}(3) &= 8 \\
m_{\mathcal{H}}(4) &= 14 \\
m_{\mathcal{H}}(n) &= \quad ? \quad 30
\end{aligned}$$