Recitation 6
Learning Theory

10-301/10-601: Introduction to Machine Learning

03/15/2024

1 Learning Theory

1.1 PAC Learning

Some Important Definitions

1. Basic notation:
   - Probability distribution (unknown): $X \sim p^*$
   - **True function** (unknown): $c^* : X \to Y$
   - **Hypothesis space** $\mathcal{H}$ and **hypothesis** $h \in \mathcal{H} : X \to Y$
   - Training dataset $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$
2. **True Error (expected risk)**

$$R(h) = P_{x \sim p^*(x)}(c^*(x) \neq h(x))$$

3. **Train Error (empirical risk)**

$$\begin{aligned}\hat{R}(h) &= P_{x \sim \mathcal{D}}(c^*(x) \neq h(x)) \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(c^*(x^{(i)}) \neq h(x^{(i)})) \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(y^{(i)} \neq h(x^{(i)}))\end{aligned}$$

The **PAC criterion** is that we produce a high accuracy hypothesis with high probability. More formally,

$$P(\forall h \in \mathcal{H}, \_\_\_\_\_\_\_\_\_\_\_\_ \leq \_\_\_\_\_) \geq \_\_\_\_\_\_\_\_$$

$$P(\forall h \in \mathcal{H}, |R(h) - \hat{R}(h)| \leq \epsilon) \geq 1 - \delta$$

Sample Complexity is the minimum number of training examples N such that the PAC criterion is satisfied for a given ϵ and δ

Sample Complexity for 4 Cases: See Figure 1. Note that

- **Realizable** means $c^* \in \mathcal{H}$
- **Agnostic** means c^* may or may not be in $\mathcal{H}$

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.
Infinite $\lvert\mathcal{H}\rvert$	**Thm. 3** $N = O(\frac{1}{\epsilon}\left[\mathrm{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\mathrm{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.

12

Figure 1: Sample Complexity for 4 Cases

The **VC dimension** of a hypothesis space $\mathcal{H}$, denoted $\mathrm{VC}(\mathcal{H})$ or $d_{VC}(\mathcal{H})$, is the maximum number of points such that there exists at least one arrangement of these points and a hypothesis $h \in \mathcal{H}$ that is consistent with any labelling of this arrangement of points.

To show that $\mathrm{VC}(\mathcal{H}) = n$:

- Show there exists a set of points of size n that $\mathcal{H}$ can shatter
- Show $\mathcal{H}$ cannot shatter any set of points of size $n+1$

Questions

1. For the following examples, write whether or not there exists a dataset with the given properties that can be shattered by a linear classifier.

   - 2 points in 1D
   - 3 points in 1D
   - 3 points in 2D
   - 4 points in 2D

   How many points can a linear boundary (with bias) classify exactly for d-Dimensions?

   - Yes
   - No
   - Yes
   - No

   $$d + 1$$

2. Consider a rectangle classifier (i.e. the classifier is uniquely defined 3 points $x_1, x_2, x_3 \in \mathbb{R}^2$ that specify 3 out of the four corners), where all points within the rectangle must equal 1 and all points outside must equal -1

   (a) Which of the configurations of 4 points in figure 2 can a rectangle shatter?

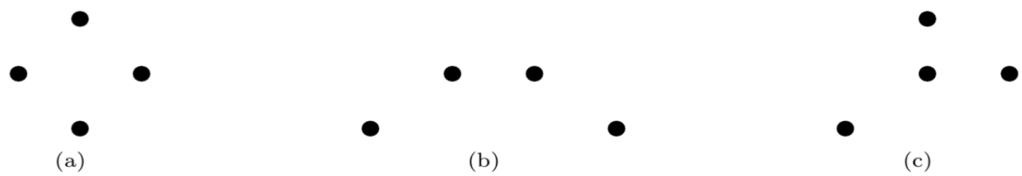


Figure 2

(a), (b), since the rectangle can be scaled and rotated it can always perfectly classify the points. (c) is not perfectly classifiable in the case that all the exterior points are positive and the interior point is negative.

   (b) What about the configurations of 5 points in figure 3?

Figure 3

None of the above. For (d), consider (from left to right) the labeling 1, 1 -1, -1, 1. For (e), same issue as (c).

3. In the below table, state in which case the sample complexity of the hypothesis falls under.

<table>
<tr><th>Problem</th><th>Hypothesis Space</th><th>Realizable/ Agnostic</th><th>Finite/ Infinite</th></tr>
<tr><td>A binary classification problem, where the data points are linearly separable</td><td>Set of all linear classifiers</td><td></td><td></td></tr>
<tr><td>Predict whether it will rain or not based on the following dataset:<table><tr><th>Temp</th><th>Humid</th><th>Wind</th><th>Rain?</th></tr><tr><td>High</td><td>Yes</td><td>Yes</td><td>Yes</td></tr><tr><td>Low</td><td>Yes</td><td>No</td><td>No</td></tr><tr><td>Low</td><td>No</td><td>Yes</td><td>Yes</td></tr><tr><td>High</td><td>No</td><td>No</td><td>Yes</td></tr></table></td><td>A decision tree with max depth 2, where each node can only split on one feature, and the features cannot be repeated along a branch</td><td></td><td></td></tr>
<tr><td>Classifying a set of real-valued points where the underlying data distribution is unknown</td><td>Set of all linear classifiers</td><td></td><td></td></tr>
<tr><td>A binary classification problem on a given set of data points, where the data is not linearly separable</td><td>K-nearest neighbour classifier with Euclidean distance as distance metric</td><td></td><td></td></tr>
</table>

	Realizable/ Agnostic	Finite/ Infinite
1	Realizable	Infinite (All possible linear classifiers)
2	Realizable (We can split the given data using a depth 2 decision tree)	Finite (There are only a finite set of decision trees that can be formed with the given constraints)
3	Agnostic (The data may or may not be linearly separable)	Infinite
4	Agnostic (The KNN classifier may or not be able to perfectly classify each point)	Finite (The hypothesis space is the set of all possible partitions of the input space into k-nearest regions - which is finite for all possible values of k)

4. Let $x_1, x_2, ..., x_n$ be n random variables that represent binary literals ($x \in \{0,1\}^n$). Let the hypothesis class $\mathcal{H}_n$ denote the conjunctions of no more than n literals in which each variable occurs at most once. Assume that $c^* \in \mathcal{H}_n$.

   Example: For $n = 4$, $(x_1 \wedge x_2 \wedge x_4), (x_1 \wedge \neg x_3) \in \mathcal{H}_4$

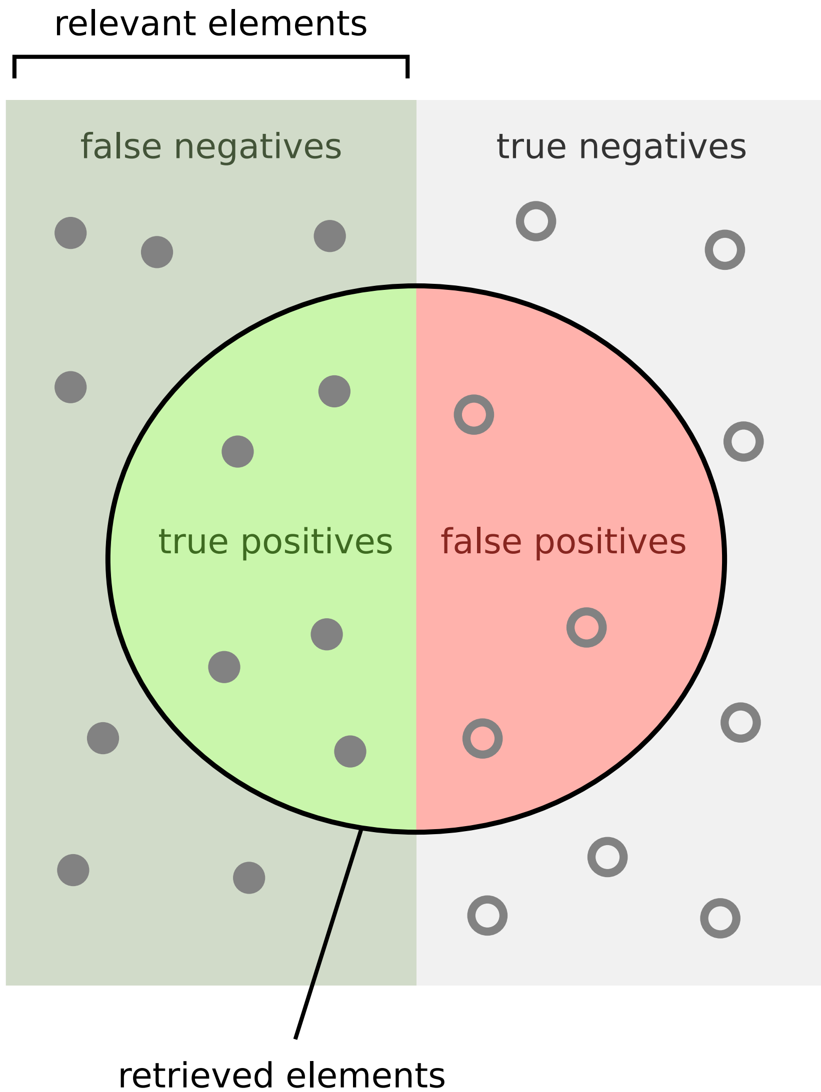
   Find the minimum number of examples required to learn $h \in \mathcal{H}_{10}$ which guarantees at least 99% accuracy with at least 98% confidence.

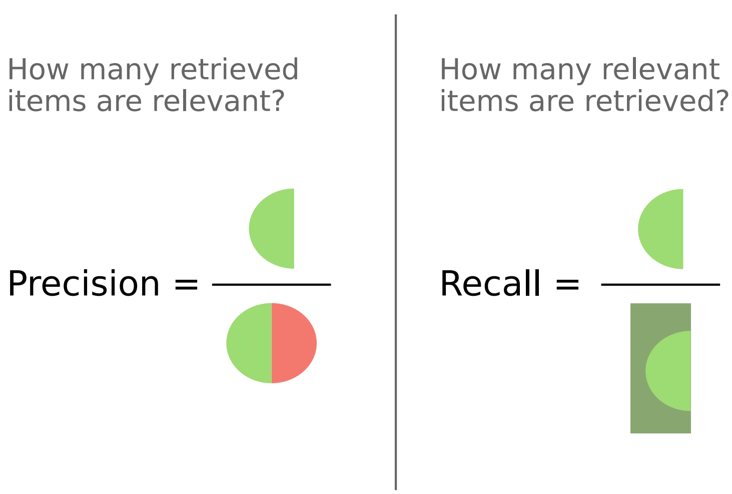
   $|H_n| = 3^n$

   $|H_{10}| = 3^{10}, \epsilon = 0.01, \delta = 0.02$

   $N(H_{10}, \epsilon, \delta) \geq \lceil \frac{1}{\epsilon}[\ln |H_{10}| + \ln \frac{1}{\delta}] \rceil = \lceil 1489.81 \rceil = 1490$

2 Precision and Recall





The following chart is known as a *confusion matrix* and helps formalize the concepts displayed above. There are 4 categories in the chart:

- *True positives*: items that are predicted positive and have actual label positive
- *False positives*: items that are predicted positive but have actual label negative
- *True negatives*: items that are predicted negative and have actual label negative
- *False negatives*: items that are predicted negative but have actual label positive

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

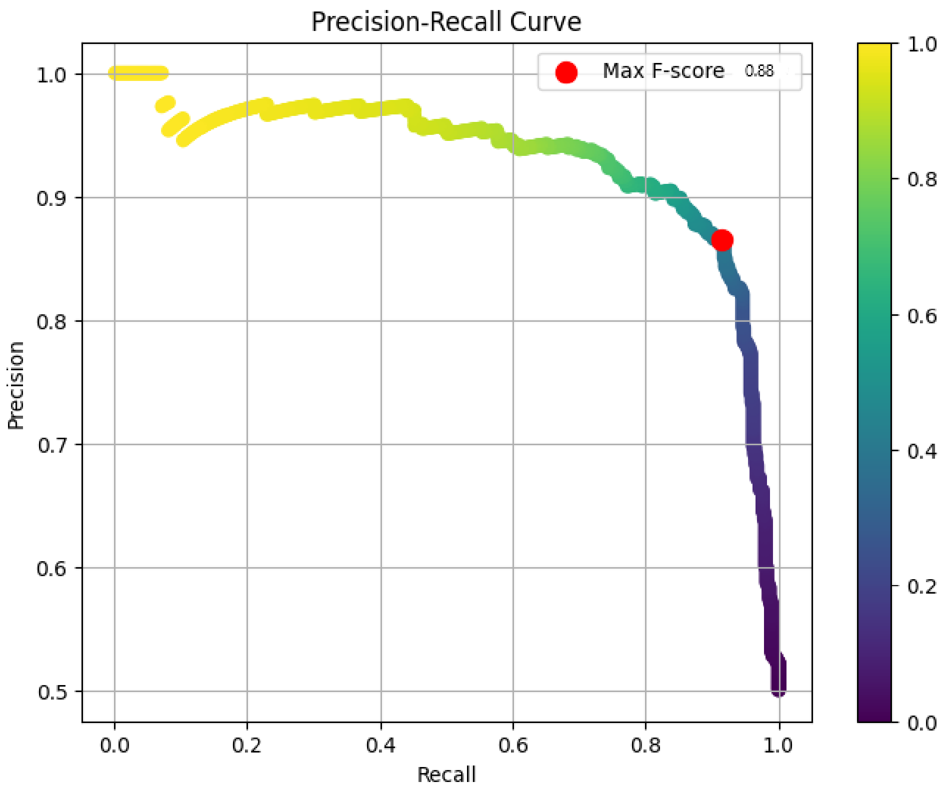
- *Type I error:* occurs when we predict a false positive (erroneously predict a positive label when the true label is negative)
- *Type II error:* occurs when we predict a false negative (erroneously predict a negative label when the true label is positive)

1. What is the formula for precision in terms of the values in the confusion matrix? What about recall? Precision = TP/(TP + FP), Recall = TP/(TP + FN)

2. The *base rate* is the proportion of items that have true label positive. What is the formula for the base rate in terms of the confusion matrix? base rate = (TP + FN) / (TP + FP + FN + TN)

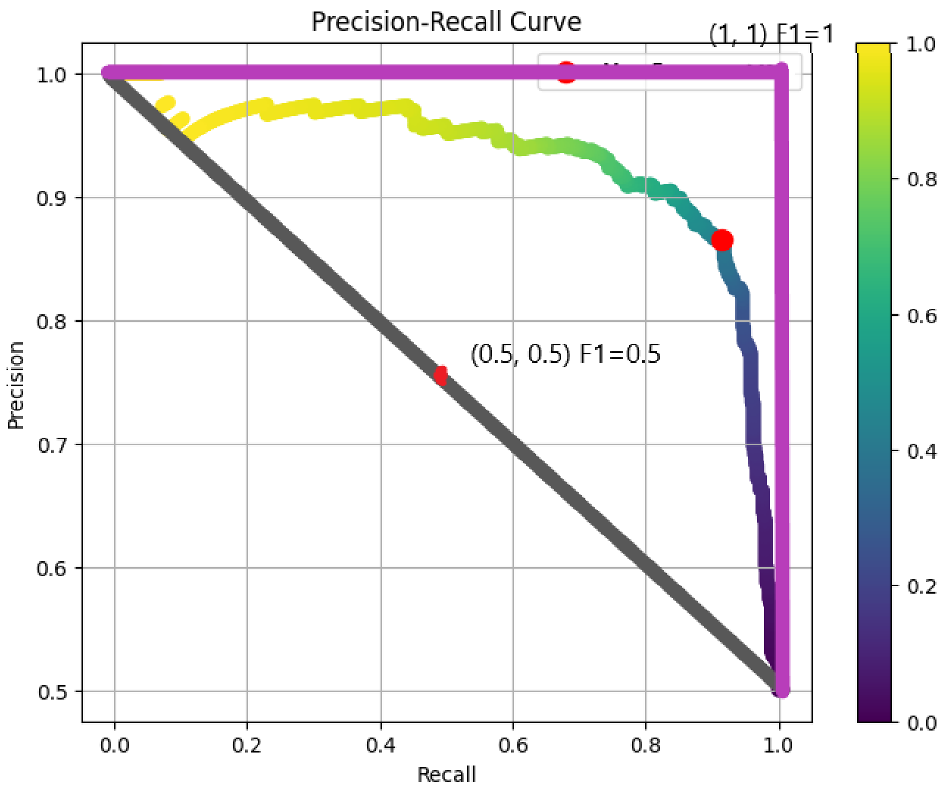
3. Suppose we predict every item to be positive. What is the precision? What is the recall? precision = base rate, recall = 1

4. The F_1 score is defined as the harmonic mean of the precision and recall: $F_1 = \frac{2}{1/P+1/R}$. The following image shows an example curve of precision and recall for a classifier when varying the threshold between the positive and negative classes. The point on the curve with highest F_1 score is marked.



Draw an example precision-recall curve for a "better" classifier than the one shown. Mark the point with the optimal F_1 score.

Draw an example precision-recall curve for a "worse" classifier than the one shown. Mark the point with the optimal F_1 score.


Precision-Recall Curve
(1, 1) F1=1
(0.5, 0.5) F1=0.5
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0