Recitation 6
PAC Learning, MLE/MAP, Societal Impacts and Fairness

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

10/24/25

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 $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$
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,

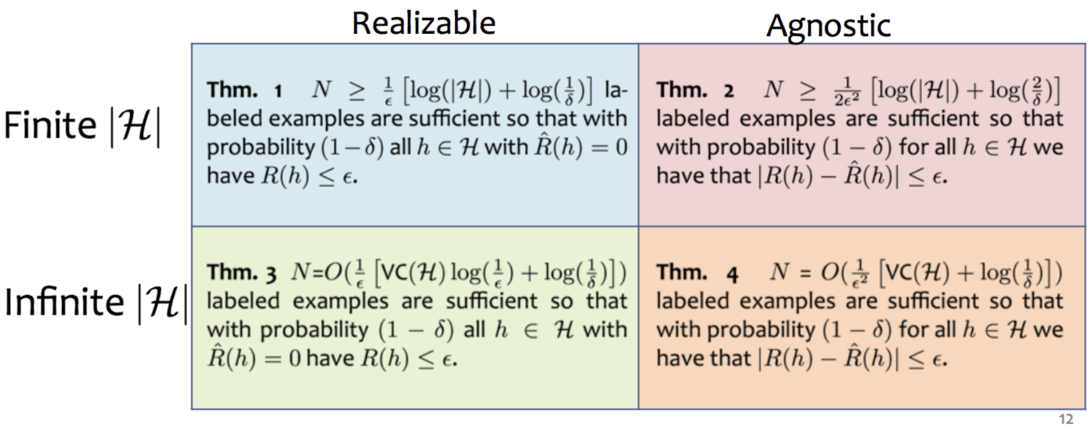
$$P(\forall h \in \mathcal{H}, \underline{\hspace{6em}} \leq \underline{\hspace{2em}}) \geq \underline{\hspace{4em}}$$

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

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

Figure 1: Sample Complexity for 4 Cases

The **VC dimension** of a hypothesis space $\mathcal{H}$, denoted 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 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

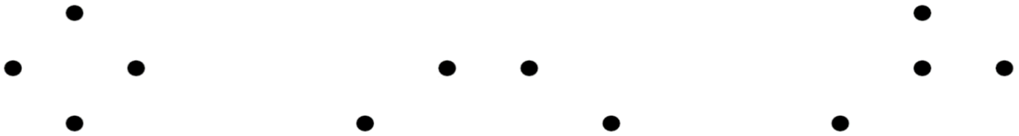
    How many points can a linear boundary (with bias) classify exactly for d-Dimensions?

    - 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?

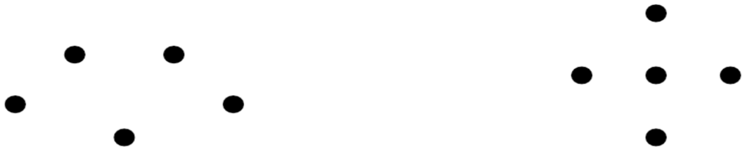


(a) (b) (c)

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?

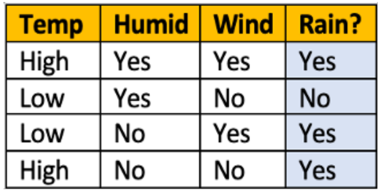


(d) (e)

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 MLE/MAP

In probabilistic learning, we are trying to learn a target probability distribution as opposed to a target function. We'll review two ways of estimating the parameters of a probability distribution: Maximum Likelihood Estimation (MLE) and Maximum a posterior (MAP) estimation.

For MLE, we have

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_{\theta} p(\mathcal{D}|\theta) \\ &= \arg\min_{\theta} -\log\left(p(\mathcal{D}|\theta)\right)\end{aligned}$$

For MAP, we have

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|\mathcal{D}) \\ &= \arg\max_{\theta} \frac{p(\mathcal{D}|\theta)p(\theta)}{\text{Normalizing Constant}} \\ &= \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta) \\ &= \arg\min_{\theta} -\log\left(p(\mathcal{D}|\theta)p(\theta)\right)\end{aligned}$$

---

1. Suppose you are a data scientist working for a customer service company that tracks the number of complaints received per day. You want to estimate the average number of complaints per day, denoted as θ. Over N days, the number of complaints observed is:

$$[x_1, x_2, \ldots, x_N]$$

Each x_i follows a Poisson distribution:

$$p(x_i \mid \theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

(a) Write the likelihood function

The likelihood function for independent Poisson observations is:

$$p(x \mid \theta) = \prod_{i=1}^{N} \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

Ignoring the factorial terms (which do not depend on θ), we get:

$$p(x \mid \theta) \propto e^{-N\theta}\theta^{\sum x_i}$$

(b) Write the log-likelihood function

Taking the natural logarithm:

$$\log p(x \mid \theta) = -N\theta + \left(\sum x_i\right)\log\theta + C$$

where C is a constant independent of θ.

(c) Take the derivative with respect to θ

$$\frac{d}{d\theta}\log p(x \mid \theta) = -N + \frac{\sum x_i}{\theta}$$

(d) Set the derivative equal to 0 and solve for $\hat{\theta}_{\text{MLE}}$.

$$-N + \frac{\sum x_i}{\theta} = 0 \Rightarrow \hat{\theta}_{\text{MLE}} = \frac{\sum x_i}{N}$$

(e) Compute the MLE estimate given $x = [3, 4, 1]$

$$\sum x_i = 3 + 4 + 1 = 8, N = 3$$
$$\hat{\theta}_{\text{MLE}} = \frac{8}{3} \approx 2.67$$

2. Suppose you are an avid Neural and Markov fan who monitors the @neuralthenarwhal Instagram account each day. You wish to determine the probability that Neural or Markov will post at any time of day. Over three days, you check Instagram and find the following number of new posts:

$$x = [3, 4, 1]$$

A fellow fan tells you that the number of posts follows a Poisson distribution:

$$p(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}$$

Furthermore, you are told that the prior distribution for θ follows a Gamma distribution:

$$\theta \sim \text{Gamma}(2, 2), \quad \text{with pdf:} \quad p(\theta) = \frac{1}{4}\theta e^{-\frac{\theta}{2}}, \quad \theta > 0.$$

(a) Derive the MAP estimator, $\hat{\theta}_{MAP}$, in terms of the data.

i. Write the log-likelihood function.

From Q1:

$$\log p(x|\theta) = -n\theta + \left(\sum x_i\right)\log\theta + C,$$

where C is a constant independent of θ.

ii. Include the prior and write the log-posterior

The prior distribution is:

$$p(\theta) \propto \theta e^{-\frac{\theta}{2}}.$$

Taking the log:

$$\log p(\theta) = \log\theta - \frac{\theta}{2} + C'.$$

Adding the log-likelihood and log-prior, the log-posterior is:

$$\log p(\theta|x) = \log p(x \mid \theta) + \log p(\theta) = -n\theta + \left(\sum x_i\right)\log\theta + \log\theta - \frac{\theta}{2} + C'',$$

which simplifies to:

$$\log p(\theta|x) = -\left(n + \frac{1}{2}\right)\theta + (\sum x_i + 1)\log\theta + C''.$$

iii. Take the derivative with respect to θ.

Differentiating the log-posterior:

$$\frac{d}{d\theta}\log p(\theta|x) = -\left(n + \frac{1}{2}\right) + \frac{\sum x_i + 1}{\theta}.$$

iv. Set the derivative to zero and solve for $\hat{\theta}_{MAP}$.

Setting the derivative to zero:

$$-\left(n + \frac{1}{2}\right) + \frac{\sum x_i + 1}{\theta} = 0.$$

Solving for θ:

$$\hat{\theta}_{MAP} = \frac{\sum x_i + 1}{n + \frac{1}{2}}.$$

(b) Compute the MAP estimate using the observed data, $x = [3, 4, 1]$.

Given $x = [3, 4, 1]$:

$$\sum x_i = 3 + 4 + 1 = 8, \quad n = 3.$$

Substituting into the formula:

$$\hat{\theta}_{MAP} = \frac{8 + 1}{3 + \frac{1}{2}} = \frac{9}{3.5} \approx 2.57.$$

Thus, the MAP estimate of θ is 2.57.

3. Compare the $\hat{\theta}_{\text{MLE}}$ and the $\hat{\theta}_{\text{MAP}}$ estimates

(a) How do the estimates differ?

- The MLE estimate $\hat{\theta}_{MLE} = 2.67$ is entirely data-driven, making it unbiased but more sensitive to fluctuations in the observed data.
- The MAP estimate $\hat{\theta}_{MAP} = 2.57$ incorporates a $\text{Gamma}(2, 2)$ prior, which slightly shrinks the estimate towards the prior belief.

Thus, MAP trades off variance for bias by incorporating prior information.

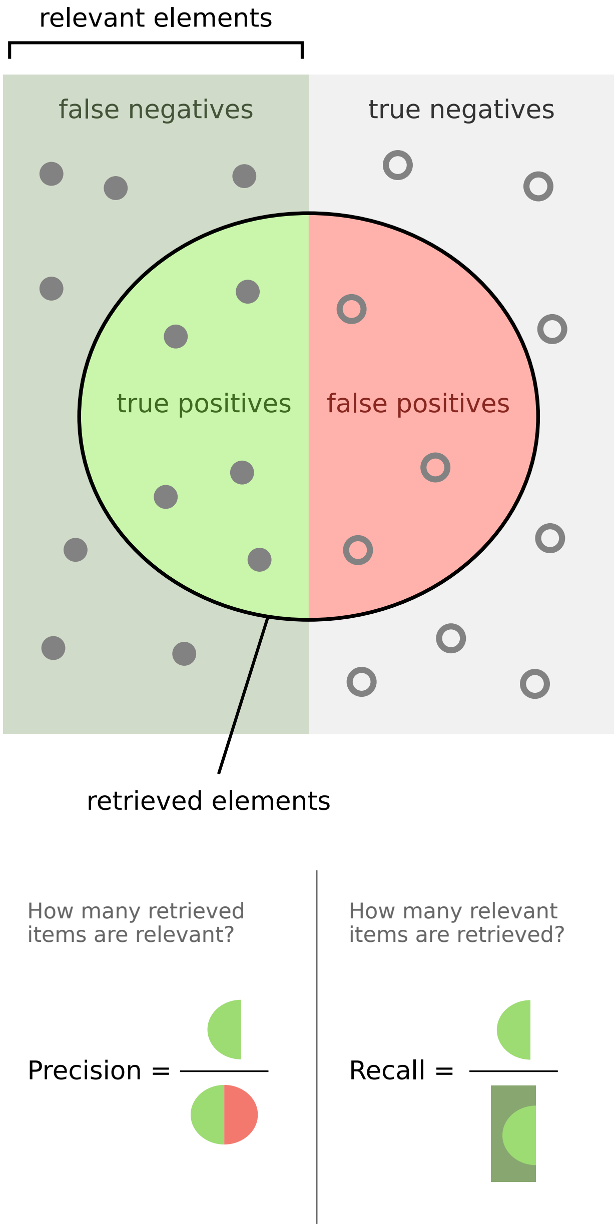
(b) Which estimate is better?

- If we trust the prior, MAP is better because it has lower variance.
- If we do not trust the prior, MLE is better because it is unbiased.

In practical settings, if we have prior knowledge, MAP is often preferred, but if we believe the data is independent and representative, MLE is reasonable.

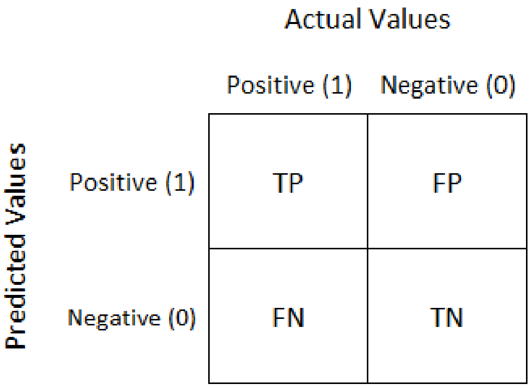
3 Societal Impacts and Fairness

3.1 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

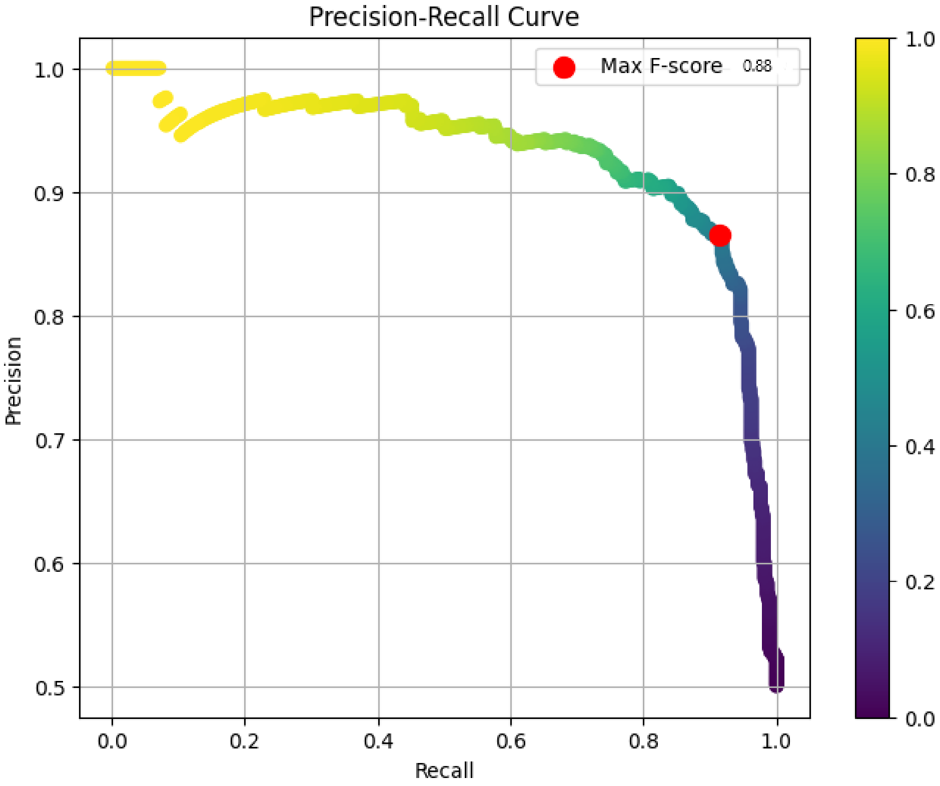
- *Type I error:* occurs when we predict a false positive (erroneously predict a positive label when the true label is negative)
- *Type II error:* occurs when we predict a false negative (erroneously predict a negative label when the true label is positive)

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)

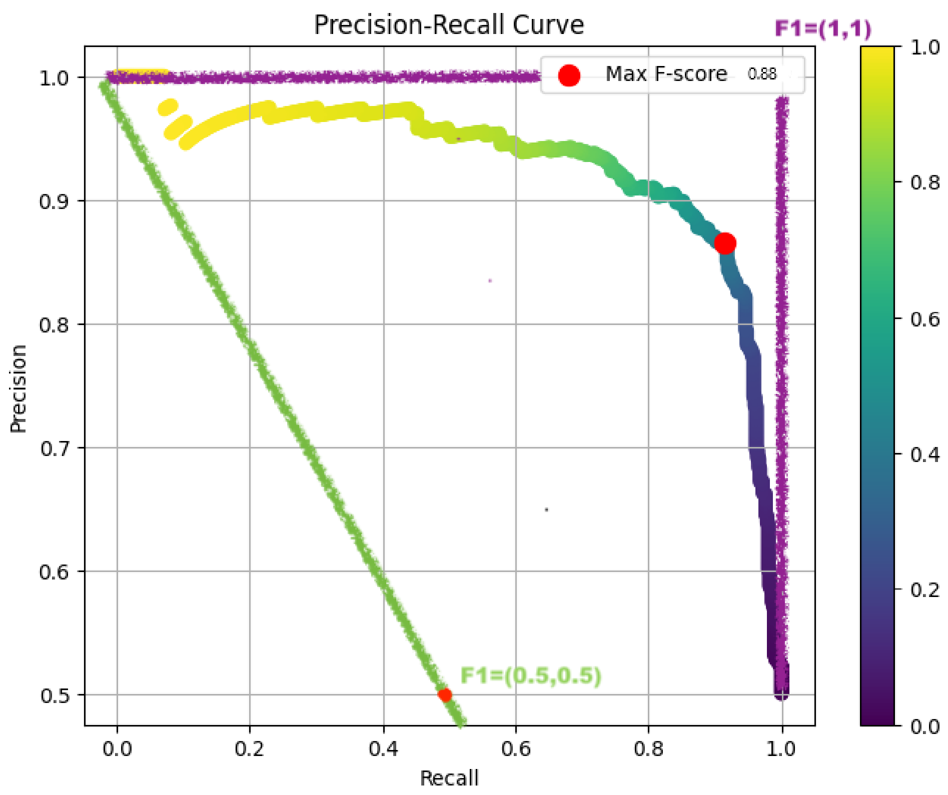
3. Suppose we predict every item to be positive. What is the precision? What is the recall? precision = base rate, recall = 1

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.



3.2 Fairness

Fairness in machine learning is a critical consideration when designing predictive models, particularly in high-stakes domains like hiring, lending, and criminal justice. Below, we outline three key definitions of fairness and their implications.

Independence (Selection Rate Parity)

- **Definition:** A model h satisfies independence if its prediction is statistically independent of the sensitive attribute A. Mathematically, $h(X, A) \perp A$.
- **Interpretation:** The proportion of accepted applicants is the same across all demographic groups.

Separation (Equality of FPR and FNR)

- **Definition:** A model satisfies separation if the false positive rate (FPR) and false negative rate (FNR) are equal across groups. Mathematically, $h(X, A) \perp A \mid Y$.
- **Interpretation:** Among truly qualified ($Y = 1$) and unqualified ($Y = 0$) applicants, the likelihood of misclassification is the same regardless of the sensitive attribute.

Sufficiency (Equality of PPV and NPV)

- **Definition:** A model satisfies sufficiency if the positive predictive value (PPV) and negative predictive value (NPV) are equal across groups. Mathematically,

$Y \perp A \mid h(X, A)$.

- **Interpretation:** Given the model prediction, the likelihood of being truly qualified ($Y = 1$) is the same across protected groups.

Each definition captures a different aspect of fairness and comes with trade-offs. The appropriate choice depends on the specific application and the ethical considerations at play. Practitioners should carefully evaluate these definitions in the context of their models and societal impact.

1. Consider the following results for a dataset with a protected attribute A with three different groups A, B, and C. Each group has exactly 200 observations in the dataset. Which of the three fairness criteria are satisfied by our model h?

Group A:

		Predicted label	
		+1	-1
True label	+1	50	30
	-1	20	100

Group B:

		Predicted label	
		+1	-1
True label	+1	60	36
	-1	24	80

Group C:

		Predicted label	
		+1	-1
True label	+1	40	24
	-1	16	120

Independence: Compute selection rates for each group:

- $SR_A = \frac{50+20}{200} = \frac{70}{200}$
- $SR_B = \frac{60+24}{200} = \frac{84}{200}$
- $SR_C = \frac{40+16}{200} = \frac{56}{200}$

Since $SR_A \neq SR_B \neq SR_C$ we conclude $h(X, A) \not\perp A$.

Separation: Compute FPR and FNR for each group:

- $FPR_A = \frac{20}{20+100} = \frac{20}{120}, FNR_A = \frac{30}{50+30} = \frac{30}{80}$
- $FPR_B = \frac{24}{24+80} = \frac{24}{104}, FNR_B = \frac{36}{60+36} = \frac{36}{96}$
- $FPR_C = \frac{16}{16+120} = \frac{16}{136} = \frac{16}{136}, FNR_C = \frac{24}{40+24} = \frac{24}{64}$

Although $FNR_A = FPR_B = FPR_C$, we see that $FPR_A \neq FPR_B \neq FPR_C$. As such, $h(X, A) \not\perp A|Y$.

Sufficiency: Compute PPV and NPV for each group:

- $PPV_A = \frac{50}{50+20} = \frac{50}{70}, NPV_A = \frac{100}{30+100} = \frac{100}{130}$
- $PPV_B = \frac{60}{60+24} = \frac{60}{84}, NPV_B = \frac{80}{36+80} = \frac{80}{116}$
- $PPV_C = \frac{40}{40+16} = \frac{40}{56}, NPV_C = \frac{120}{24+120} = \frac{120}{144}$

Although $PPV_A = PPV_B = PPV_C$, the NPVs across groups are not equal. Hence, $Y \not\perp A|h(X, A)$

As illustrated above, it is difficult to achieve all three fairness criteria at once. In general, any pair of these criteria are mutually exclusive in almost all situations!