Foundation of Data Science
Lecture 6, Module 1
Fall 2022

Rumi Chunara, PhD

Evaluation metrics


Ceci n'est pas une pipe.

What is a statistical model?

- A model is a representation of an idea, an object, a process or a system that is used to describe and explain phenomena that cannot be experienced directly
  - Stands for something
  - Describes patterns in data in reduced dimensions

Reminder

You will never build the *perfect* model… but we can always have the *best possible* model.

So far we have discussed the following design options:

[Data, Algorithm, Feature Set , Hyper-parameters (complexity)]

We also need to choose an evaluation metric!

The right metric depends on your goals

- **Classification** – Is this email spam or not? Is this number a ‘1’ or a ‘7’?

- **Regression** - What is the price of a house based on its features (size, neighborhood, year it was built, etc?)

- **Density Estimation** – What is the probability that this transaction is fraud? What is the expected spending of a new credit card customer? We’ll discuss this in the future

Metrics for these Goals

Classification

Focus today!

Recall (RCL)
Precision (PRE)
F-Score (FSC)
Accuracy (ACC)
Area under the Receiver Operator Curve (AUC)

Regression

Focus today!

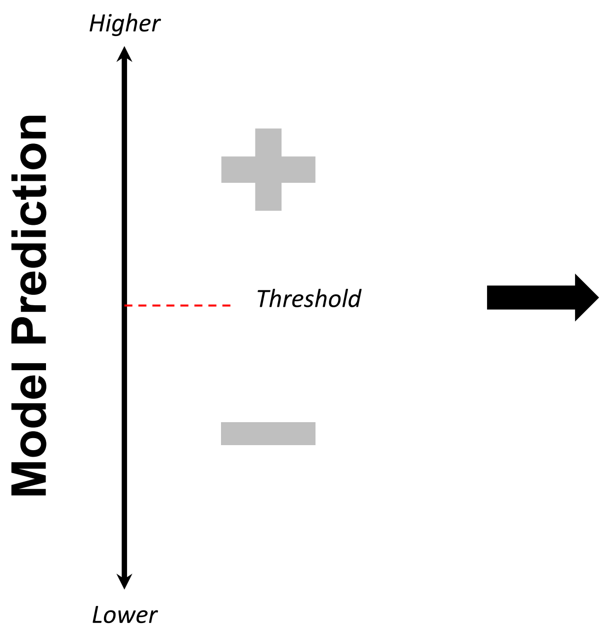
Mean Absolute Error (MAE)
Mean Squared Error (MSE)
Coefficient of Determination (R-squared)

Classification Metrics

Confusion Matrix

Many of the metrics we use are derived from the confusion matrix. For binary classification we assume there exists some real valued function f(x) and a decision threshold δ.

$$\hat{Y} = I(f(x) > \delta)$$



		actual value		
		p	***n***	**total**
prediction outcome	***p'***	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

Classification Metrics

We can derive many classification metrics from the confusion matrix.

		actual value		
		p	n	total
prediction outcome	p'	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

Terminology and derivations from a confusion matrix

true positive (TP)
eqv. with hit

true negative (TN)
eqv. with correct rejection

false positive (FP)
eqv. with false alarm, Type I error

false negative (FN)
eqv. with miss, Type II error

sensitivity or true positive rate (TPR)
eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

false positive rate (FPR)
eqv. with fall-out

$$FPR = FP/N = FP/(FP + TN)$$

accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

specificity (SPC) or True Negative Rate

$$SPC = TN/N = TN/(FP + TN) = 1 - FPR$$

positive predictive value (PPV)
eqv. with precision

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

false discovery rate (FDR)

$$FDR = FP/(FP + TP)$$

Matthews correlation coefficient (MCC)

$$MCC = (TP * TN - FP * FN)/\sqrt{PNP'N'}$$

F1 score

$$F1 = 2TP/(P + P') = 2TP/(2TP + FP + FN)$$

Source: Fawcett (2006).

Source:
http://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_curve

Recall

Measure of how much relevant information the system has extracted (coverage of system).

Basic idea:

$$\text{Recall} = \frac{\text{\# of correct positive labels given by system}}{\text{total \# of possible positive labels}}$$

Recall

Measure of how much relevant information the system has extracted (coverage of system).

Exact definition:

Recall = 1 if no possible correct answers

else:

$$\frac{\text{\# of correct positive labels given by system}}{\text{total \# of possible positive labels}}$$

Precision

Measure of how much of the information the system returned is correct (accuracy).

Basic idea:

$$\text{Precision} = \frac{\text{\# of correct positive labels given by system}}{\text{\# positive labels given by system}}$$

Slide from Butt/Jurafsky/Martin

Precision

Measure of how much of the information the system returned is correct (accuracy).

Exact definition:

Precision = 1 if no answers given by system

else:

$$\frac{\text{\# of correct positive labels given by system}}{\text{\# positive labels given by system}}$$

Evaluation

Every system, algorithm or theory should be **evaluated**, i.e. its output should be compared to the **gold standard** (i.e. the ideal output). Suppose we try to find scientists…

Algorithm output:
O = {Einstein, Bohr, Planck, Heisenberg, Obama}

Gold standard:
G = {Einstein, Bohr, Planck, Heisenberg}

Precision:
What proportion of the output is correct?

$$\frac{|O \wedge G|}{|O|}$$

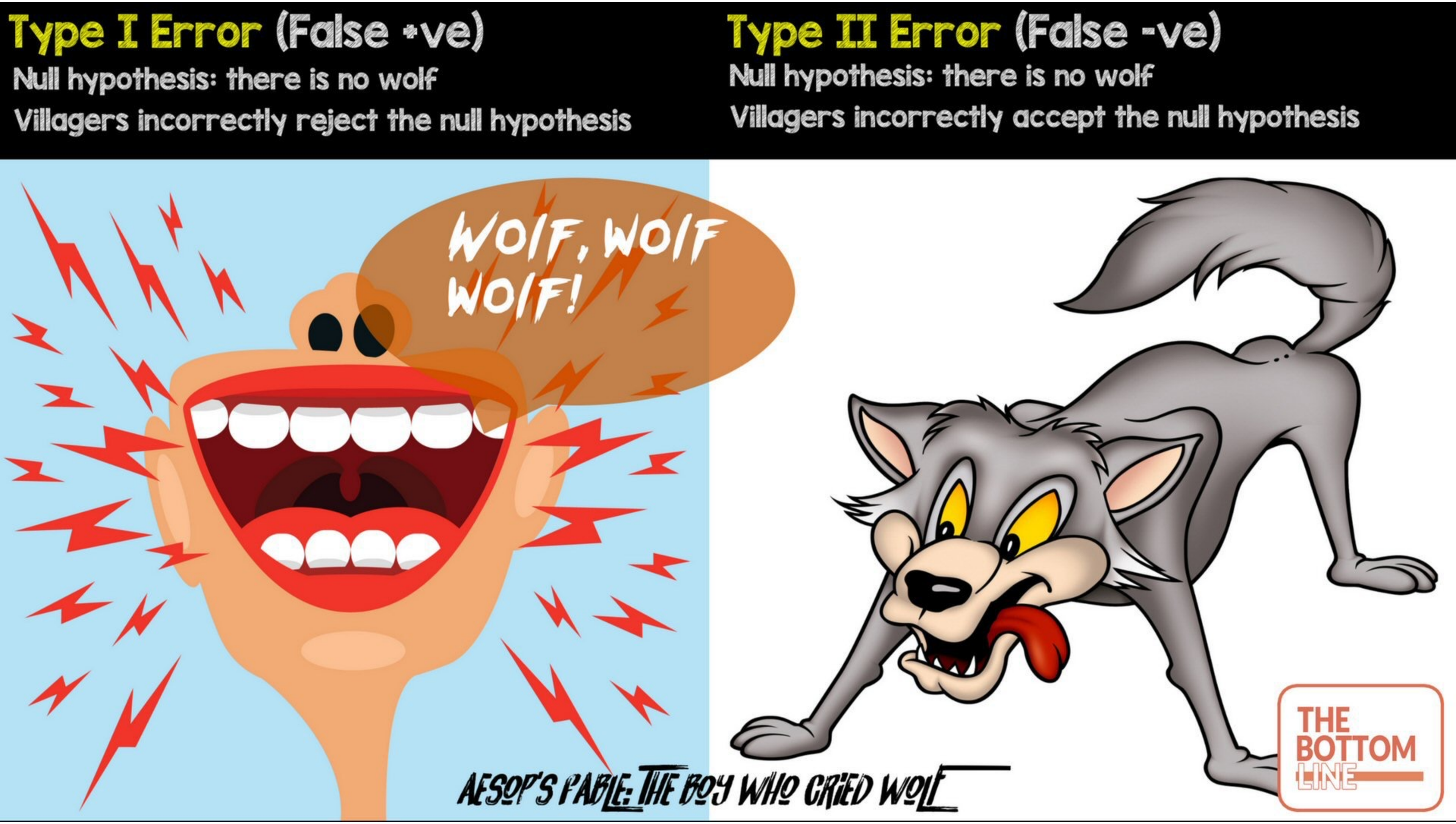
Recall:
What proportion of the gold standard did we get?

$$\frac{|O \wedge G|}{|G|}$$

Slide modified from Suchanek

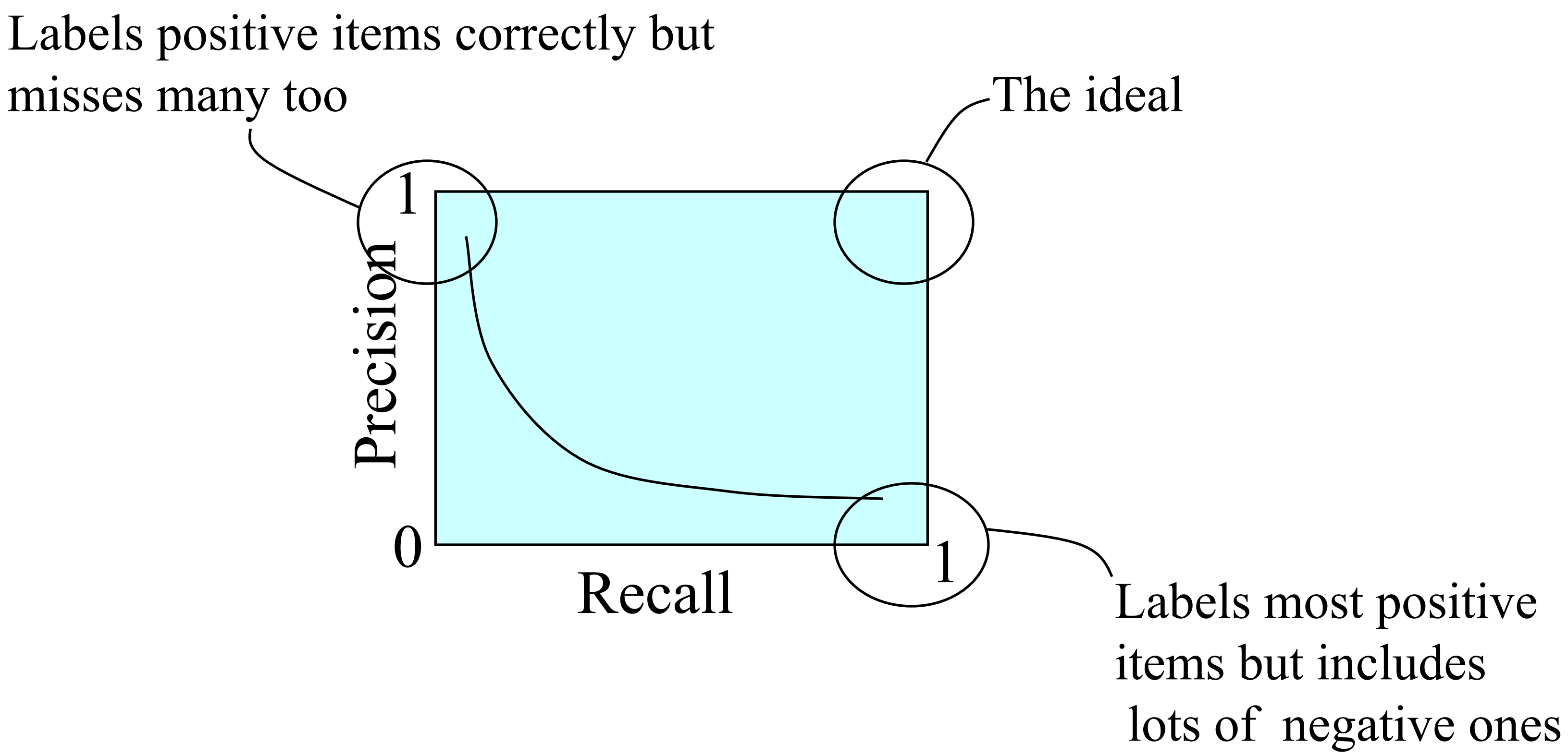
Types of Errors

- False Positives
  - The system predicted **TRUE** but the value was **FALSE**
  - aka "False Alarms" or Type I error
- False Negatives
  - The system predicted **FALSE** but the value was **TRUE**
  - aka "Misses" or Type II error



https://twitter.com/mathematicsprof/status/1302109456294174722

Trade-off between Recall and Precision



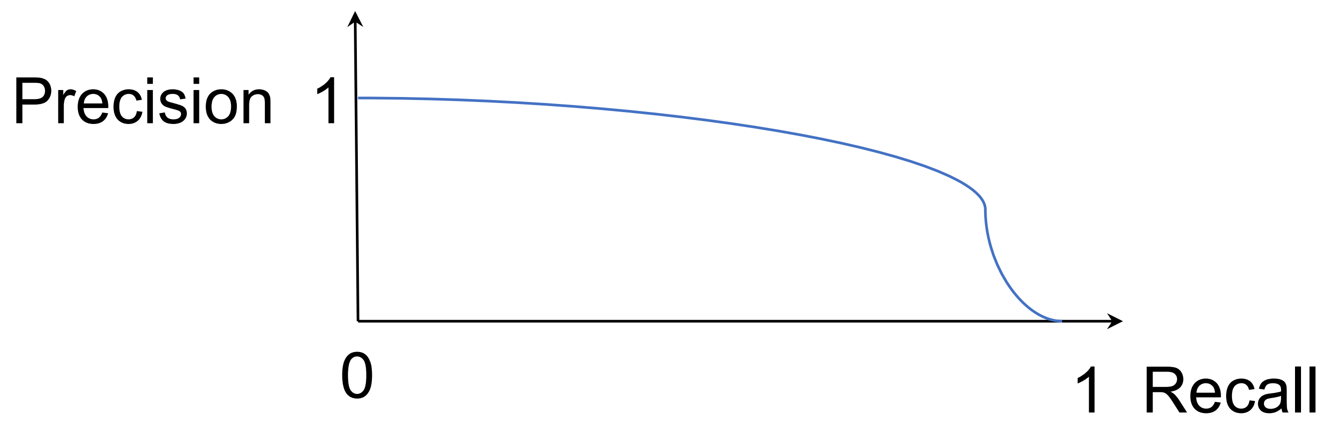
Precision/Recall

- You can get high recall (but low precision) by returning positive labels for all items!

- In a good system, precision often decreases as the recall increases
  - This is not a theorem, but a result with strong empirical confirmation

F1- Measure

You can't get it all...



The F1-measure combines precision and recall as the harmonic mean:

F1 = (2 * precision * recall) / (precision + recall)

F-measure

Precision and Recall stand in opposition to one another. As precision goes up, recall usually goes down (and vice versa).

The F-measure combines the two values.

$$\text{F-measure} = \frac{(ß^2+1)PR}{ß^{2*}P+R}$$

- When ß = 1, precision and recall are weighted equally (same as F1).
- When ß is > 1, precision is favored.
- When ß is < 1, recall is favored.

F: Example

	positive	negative	total
Labeled positive	20	40	60
labeled negative	60	1,000,000	1,000,060
total	80	1,000,040	1,000,120

F: Example

	positive	negative	total
Labeled positive	20	40	60
Labeled negative	60	1,000,000	1,000,060
total	80	1,000,040	1,000,120

- P = 20/(20 + 40) = 1/3
- R = 20/(20 + 60) = 1/4

$$F_1 = 2\frac{1}{\frac{1}{\frac{1}{3}}+\frac{1}{\frac{1}{4}}} = 2/7$$

F1 value lies somewhere in the middle!

Should we use accuracy instead?

- Assume that an algorithm classifies each item as positive or negative

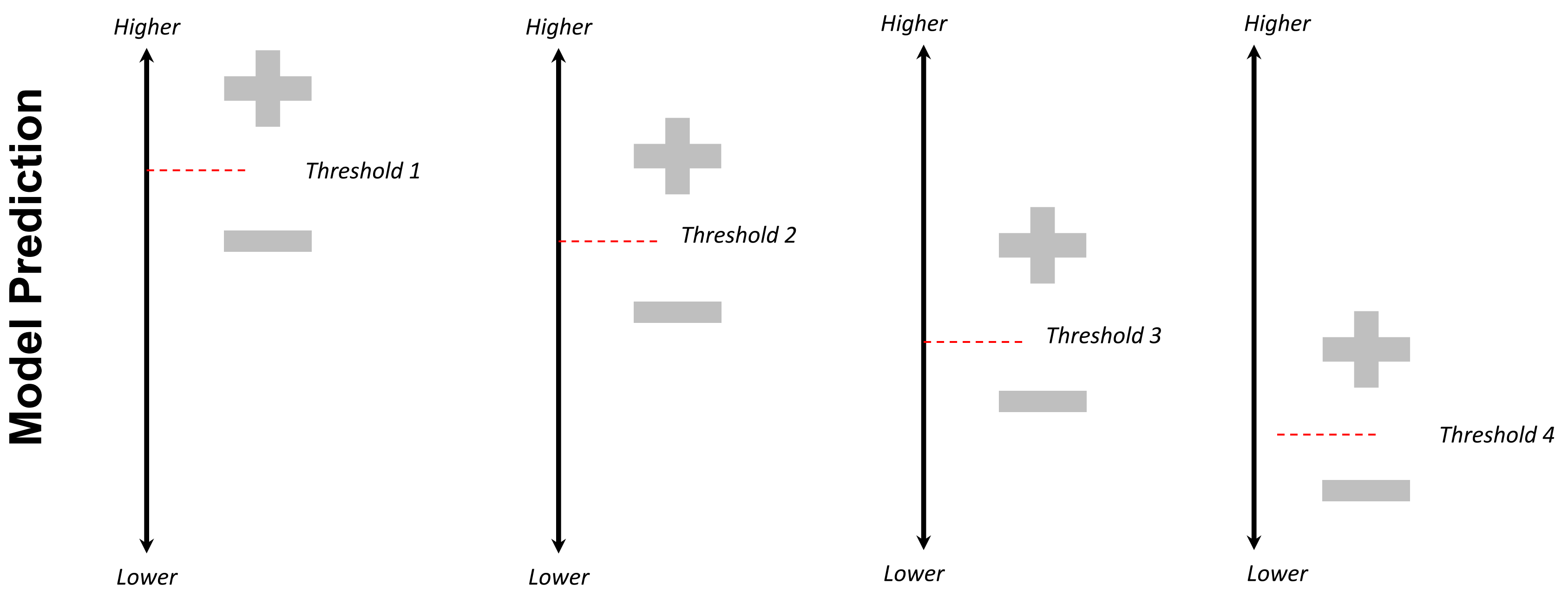
- The accuracy is the fraction of these labels that are correctly classified

$$\text{Accuracy} = (tp + tn) / (tp + fp + fn + tn)$$

- Accuracy is a commonly used evaluation metric in machine learning
  - What are the limitations of this evaluation metric?

Towards a Ranking Metric

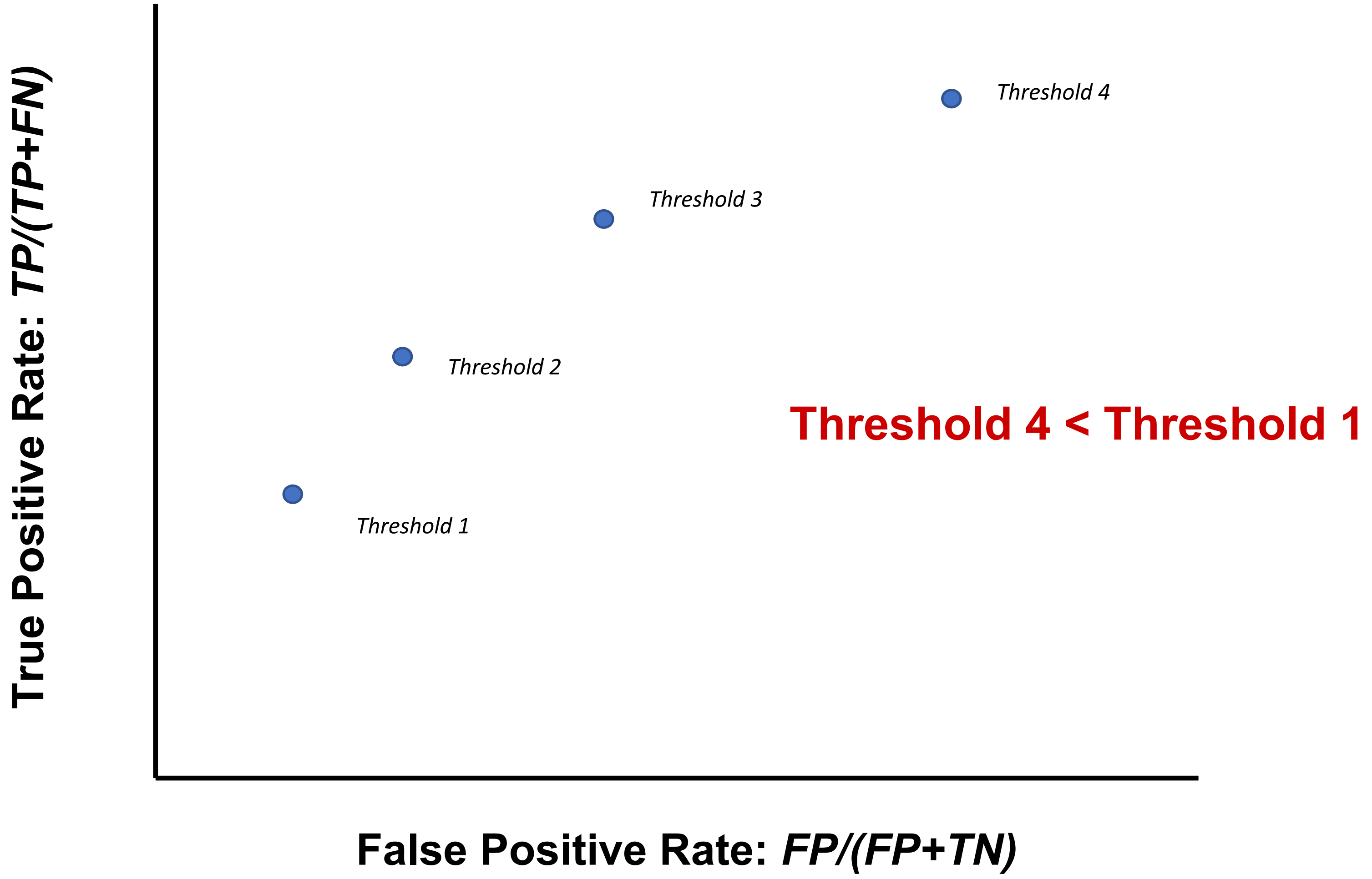
Classification metrics depend on choosing a single threshold. But what if you don't know or need the threshold?



For each threshold we will get different recall, precision, and accuracy.
We want an evaluation method that considers the trade-off on these metrics when using different thresholds.

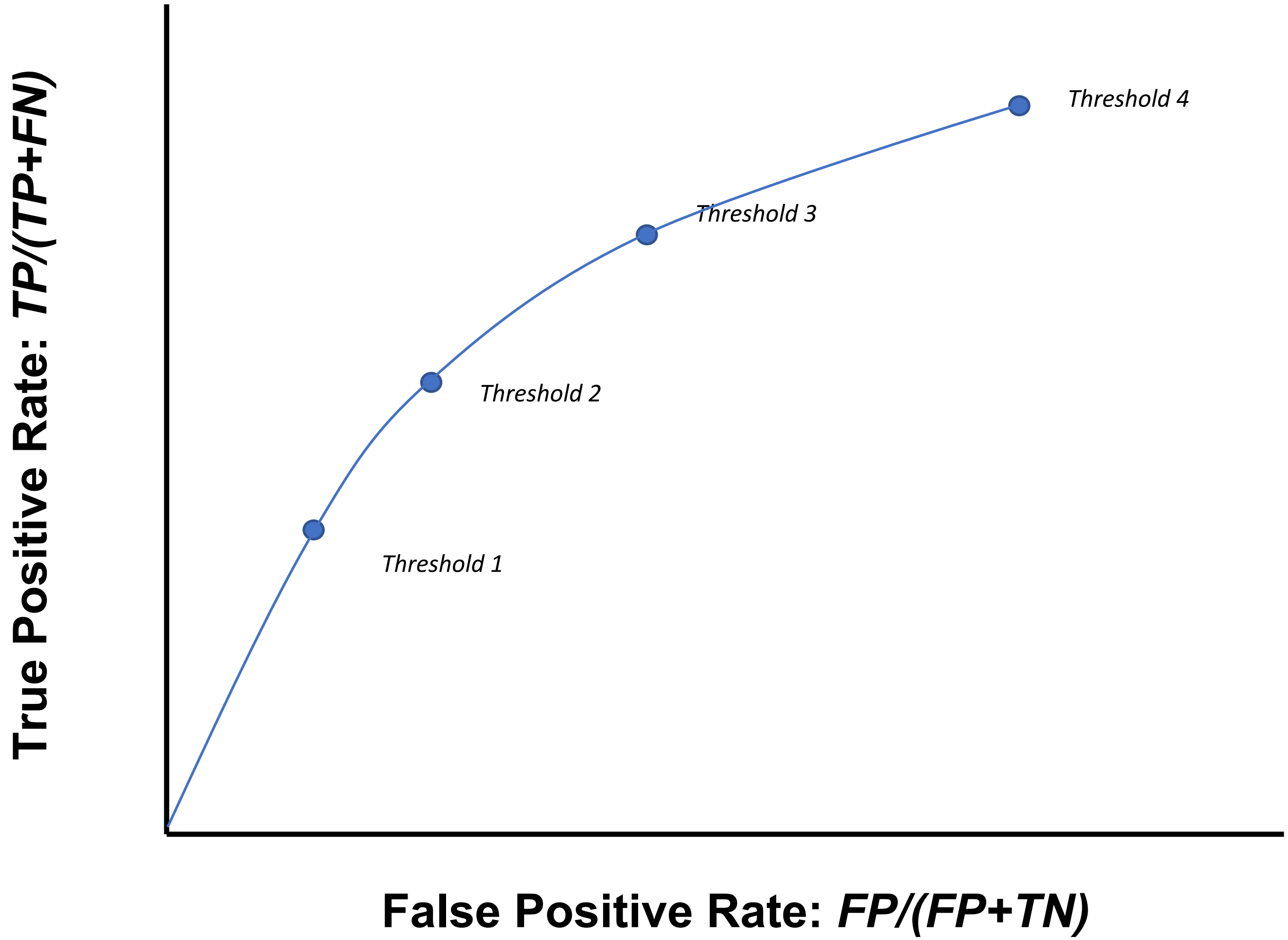
More on The Thresholding Trade-Off

Each threshold we choose creates a trade-off between false positive rate and true positive rate.



The ROC Curve

If we consider every threshold and plot the trade-off, we arrive at the ROC curve.

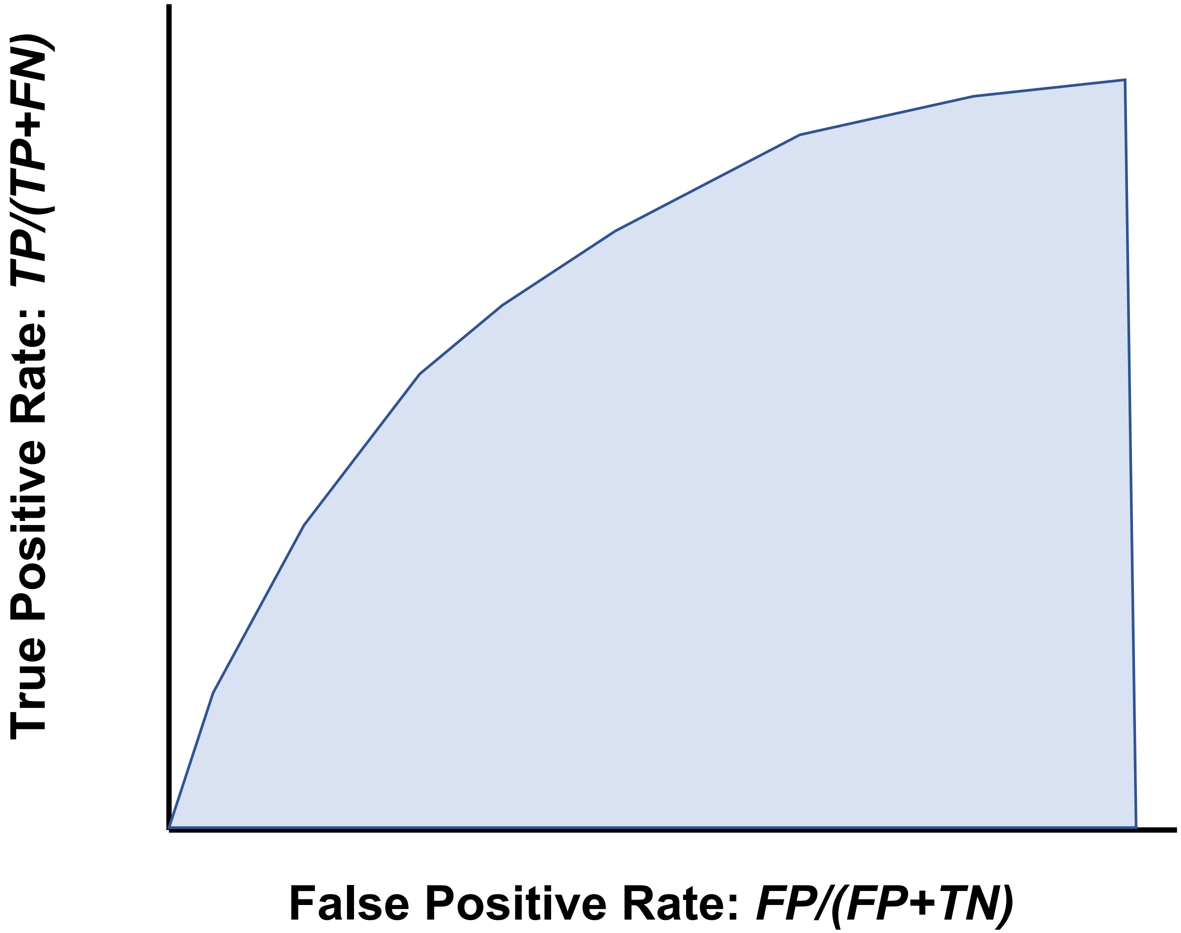


Note that True Positive Rate = Recall

Scope: binary classification

The Area Under the ROC Curve

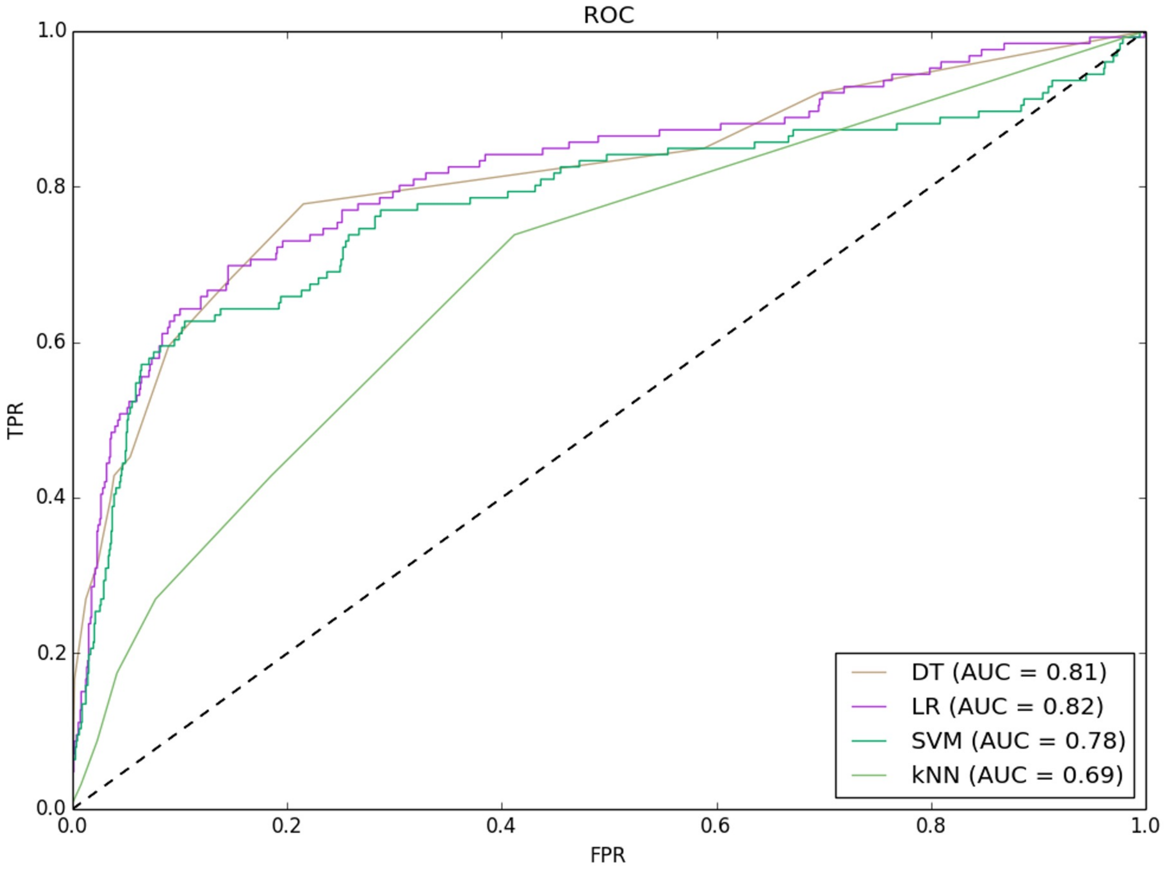
The area under this curve gives a comprehensive summary of how well your classifier performs.



Comparing AUCs

We built 4 different classifiers using an ads dataset. We can compare the models using ROC analysis.

- A universally better model has higher TPR at all FPR (LR > kNN)
- Some models overlap. Better model depends on whether you value TPR or FPR more (DT is best where FPR < 0.05)



Fun AUC Facts

- **Nice interpretation**: AUC gives the probability that a positive instance will have a higher score than a negative instance (equivalent to Mann-Whitney U statistic)

- **Scale invariant:** AUC measures how well predictions are ranked, rather than their absolute values.

- **Is nicely bounded:** AUC scores range from [0,1], where 1 is a perfect classifier and 0 is a perfectly wrong classifier. A random classifier has an exact score of 0.5.

- **Classification threshold invariant:** AUC measures the quality of the model's predictions irrespective of what classification threshold is chosen.

Regression Metrics

Mean Absolute Error (MAE)

MAE is a measure of errors between a prediction and the actual outcome.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

- MAE uses the same scale as the data being measured (scale-dependent accuracy measure)
- It cannot be used with a series of points in different scales
- Commonly used in time series analysis
- Relatively robust to the presence of outliers

Mean Squared Error (MSE)

MSE measures the average of the squares of the error between a prediction and the outcome.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- It is always non-negative, making certain mathematical analyses easier
- It is a differentiable function that makes it easy to perform mathematical operations in comparison to a non-differentiable function like MAE

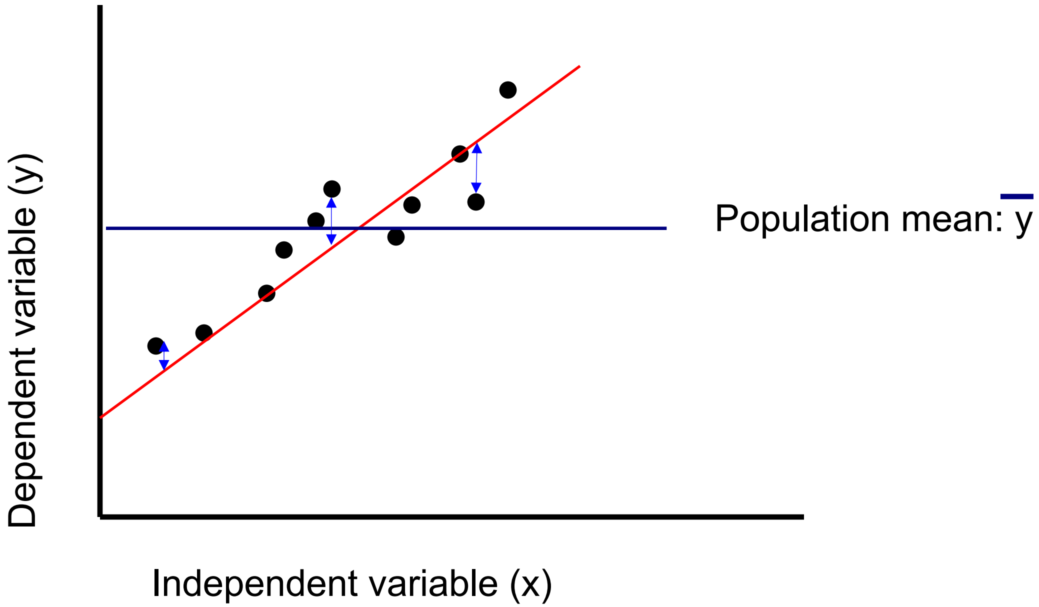
Root Mean Squared Error (RMSE)

RMSE is the square root of MSE.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

- RMSE is more popular than MSE (it yields smaller values that are often easier to interpret)
- It is also non-negative and differentiable

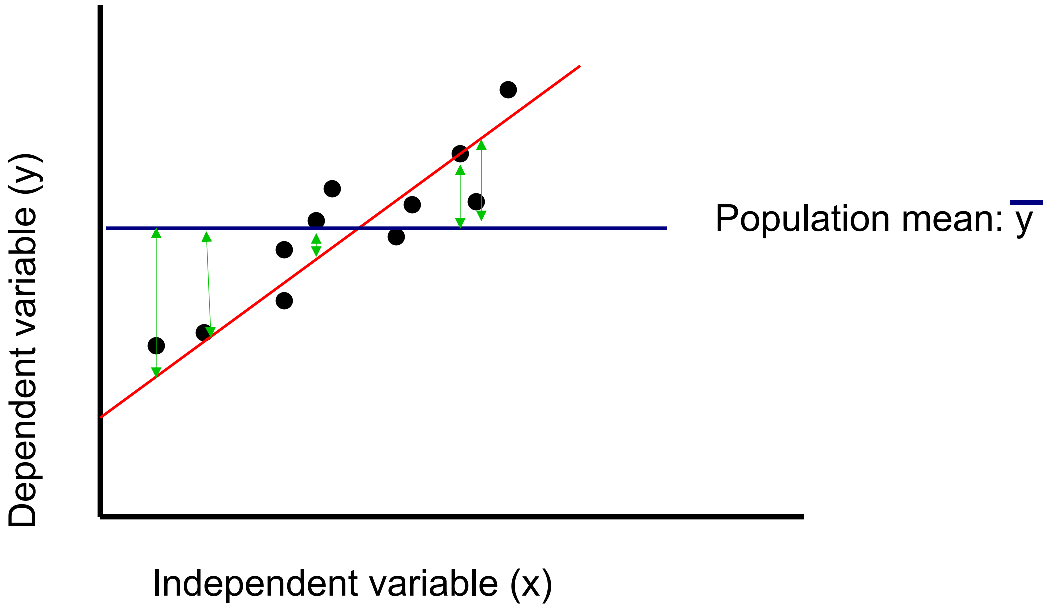
Linear regression model variation



The Sum of Squares Regression (SSres) is the sum of the squared differences between the prediction for each observation and the population mean.

The responses y_i correspond to different values of the explanatory variable x and will differ because of that. The fitted values y_i estimates the mean response for the specific x_i. The differences $y_i - \hat{y}$ reflect the variation in mean response due to differences in the x_i.

Linear regression model variation



The Total sum of squares (SStot) is the sum of the squared differences between the prediction for each observation and the mean value $(y_i - \bar{y})$.

Coefficient of determination (R-squared)

The coefficient of determination (**R-squared**) is the proportion of the variation in the dependent variable that is predictable from the independent variable.

$$SS_{\text{res}} = \sum_i (y_i - \hat{y}_i) = \sum_i e_i^2$$

Residual sum of squares

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

Total sum of squares

$$\boxed{R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}}$$

- R-squared is a scale-free score, and the maximum value is 1 (the larger the better)
- Negative R-squared values can occur when predictions are worse than random (SSres > SStot)

To think about…

- Where is the data from? Was it collected for the purpose you are using it? Are there any limitations to the data due to this?

- For your project, what are the appropriate evaluation metric(s)?

- Are there any important subgroups in the data? How does performance compare across subgroups?

- Who are your stakehodlers? What are important results to communicate?