Module 25 – Precision and Recall

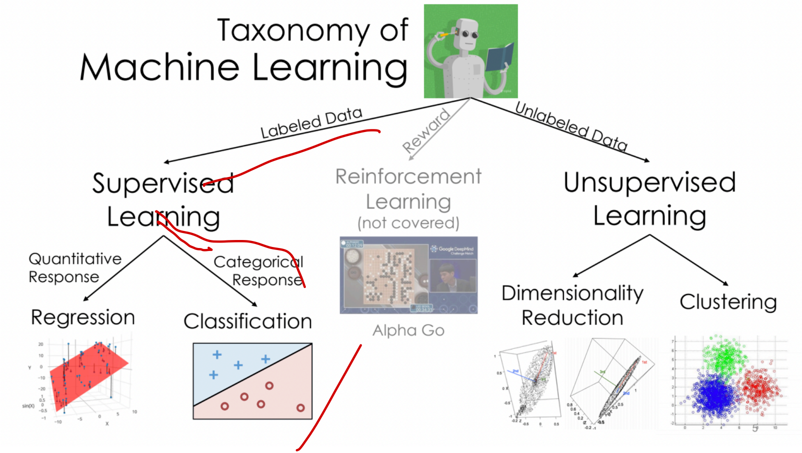
DSC 40A, Summer 2023

Agenda

- Measuring quality of classification

Measuring quality of classification

Taxonomy of machine learning



1

[1] taken from Joseph Gonzalez @ UC Berkeley

Classification problems

- In the classification problem, we make predictions based on data (called **training data**) for which we know the value of the **categorical** response variable.
- Example classification problems:
  - Deciding whether a patient has kidney disease.
  - Identifying handwritten digits.
  - Determining whether an avocado is ripe.
  - Predicting whether credit card activity is fraudulent.

Assessing the quality of a classifier

- SVM
- K-nn classifier

- Naive Bayes is one classification algorithm, or **classifier**, but there are many others.
  - logistic
  - deep learning (any architecture)
  - GBM (Xgboost and Lightgbm)
  - Random forest
  - decision tree

- Is Naive Bayes any good? How do we measure how good of a job a classifier does? differentiate between..

evaluation metrics

cost/empirical risk,

Discussion Question

Think back to regression (supervised learning with a quantitative response variable). How did we measure the quality of our predictions? Can we adopt a similar strategy?

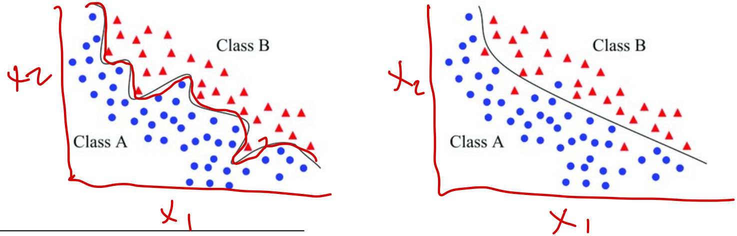
Unseen data

- A natural way to measure the quality of our classifications is to see how often we predict the right category.

  aka accuracy

- We want to make good predictions on **unseen data**. So we'll measure how often we classify examples correctly for a new set of **test data**.

- This provides us an indication about whether or not we are overfitting, and also helps us estimate general out of sample performance.



2

[2]Park, Hochong, Son, Joo-Hiuk. (2021). Machine Learning Techniques for

Accuracy ← evaluation metric, not cost or loss

- Classification **accuracy** is the proportion of examples in the test set that are correctly classified.
- Accuracy is measured on a 0 to 1 scale.

Accuracy

- We can think of accuracy as an estimate for the probability of making a correct classification on an unseen example.
- Parameter:
$$P(\text{successful classification})$$
- Estimate:
$$\text{accuracy} = \frac{\text{\# correctly classified examples in test set}}{\text{size of test set}}$$

Imbalanced classes

Alagille syndrome is a rare genetic condition that affects 1 in 40,000 people. We want to classify people as having this condition (**unhealthy**) or not having this condition (**healthy**).

1

0

Discussion Question

Consider a classifier that classifies everyone as **healthy**.

prediction = 0 for all cases

1. What is the accuracy of this classifier? $\frac{39{,}999}{40000}$
2. What are the ethical repercussions of using this classifier?

High accuracy is not enough

- We want to avoid overdiagnosis (telling someone they have the condition when they don’t).

- We also want to avoid underdiagnosis (telling someone they’re healthy when they’re not).

- It’s easy to avoid either one of these. It’s hard to avoid both of these simultaneously, yet a good classifier should do exactly that.

Different types of errors

where "unhealthy" = positive, i.e. have disease

1

Confusion matrix

Ground truth

predicted

	1 Actually unhealthy	0 Actually **healthy**
1 Classified as **unhealthy**	True Positive	False Positive Type I error
0 Classified as **healthy**	False Negative / Type 2 error	True Negative

Avoid overdiagnosis

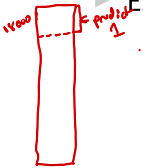
	Actually **unhealthy**	Actually **healthy**
Classified as **unhealthy**	True positive	False positive
Classified as **healthy**	False negative	True negative

- How often does our prediction of the condition mean a person actually has the condition?
- Parameter:

$$\frac{TP}{TP+FP}$$

$$P(\text{actually } \textbf{unhealthy} \mid \text{classified as } \textbf{unhealthy})$$

- Estimate:



$$\textbf{precision} = \frac{\text{\# people in test set } \textbf{correctly} \text{ classified as } \textbf{unhealthy}}{\text{\# people in test set classified as } \textbf{unhealthy}}$$

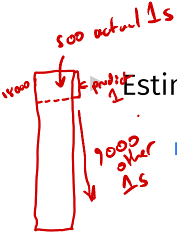
Avoid underdiagnosis

	Actually **unhealthy**	Actually **healthy**
Classified as **unhealthy**	True positive	False positive
Classified as **healthy**	False negative	True negative

- How often do we identify those that actually have the condition?

- Parameter:

$$P(\text{classified as } \textbf{unhealthy} \,|\, \text{actually } \textbf{unhealthy})$$



- Estimate:

$$\textbf{recall} = \frac{\text{\# people in test set } \textbf{correctly} \text{ classified as } \textbf{unhealthy}}{\text{\# } \textbf{unhealthy} \text{ people in test set}}$$

Precision vs. recall

	Actually **unhealthy**	Actually **healthy**
Classified as **unhealthy**	True positive	False positive
Classified as **healthy**	False negative	True negative

- Precision:

$$\textbf{precision} = \frac{\text{\# people in test set } \textbf{correctly} \text{ classified as } \textbf{unhealthy}}{\text{\# people in test set classified as } \textbf{unhealthy}}$$

$$= \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

- Recall:

$$\textbf{recall} = \frac{\text{\# people in test set } \textbf{correctly} \text{ classified as } \textbf{unhealthy}}{\text{\# } \textbf{unhealthy} \text{ people in test set}}$$

$$= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

ideally, we want both to be 1.0

Precision vs. recall

Discussion Question

Consider a marketing model where you are optimizing who to show targeted ads to. In this case, our labels are "converted" vs "not converted." In our training sample, we only have .0001 positive records (i.e. a severe class imbalance).

Assume sample is 5 mil

1. Suppose you select the top 500 ranked records as an arbitrary decision threshold. What are some reasonable values that you might expect for precision and recall?

2. If you were to increase the selection to the top 500,000, would you expect precision to go up or down? What about recall?

Combining precision and recall

- We want high precision and high recall, but it's hard to have both.

- Let's combine them into a single measurement.

- Does the average of precision and recall work well?

$$\frac{P + R}{2}$$

- Compare:
  - 00001
  - Classifier A ($P = 0$; $R = 1$) $= 0.5$
  - Classifier B ($P = 0.5$, $R = 0.6$) $= 0.55$

Combining precision and recall

- **Key insight:** Two moderate values are better than two extremes. Use the product, which shrinks when either term in the product is small.

- New way of combining precision and recall: **F-score**

$$\frac{2PR}{P + R}$$

- Compare:
  - Classifier A ($P = 0$, $R = 1$) $= \frac{2 \cdot 0 \cdot 1}{0 + 1} = 0$
  - Classifier B ($P = 0.5$, $R = 0.6$) $= \frac{2 \cdot (.5) \cdot (.6)}{0.5 + 0.6} = \frac{6}{11}$

F-score

- The **F-score** combines the precision and recall of a classifier in a single measurement.

$$\frac{2PR}{P + R}$$

$P = 1$
$R = 1$

- Higher F-score $\Rightarrow$ better classifier.

Discussion Question

What would be the F-score of a "perfect classifier"?

1

Summary

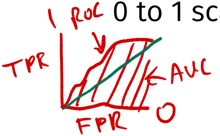
Summary

- Accuracy is a simple way of measuring the quality of a classifier, but it can be misleading when classes are imbalanced.

sensitivity

- Precision and recall are two other ways of measuring the quality of a classifier, but they can be hard to achieve simultaneously.

- The F-score combines precision and recall into a single measurement that assesses the quality of a classifier on a 0 to 1 scale.



$$\frac{2PR}{P+R}$$