Evaluating Models

Brian Kim

Motivation

Suppose we have a machine learning model (e.g. decision trees) that we fit on a **training set**, and we need to now evaluate it on the **test set**. How should we do this?

Example

Suppose the following cars were in the test set.

Horsepower	Miles Per Gallon	Outcome
200	15	Success
100	15	Failure
150	25	Success
122	31	Failure
110	12	Failure
304	11	Failure
283	15	Failure

Example

Using the training set, we built a model, and produced the following predictions

Horsepower	Miles Per Gallon	Outcome	Prediction
200	15	Success	Success
100	15	Failure	Failure
123	25	Success	Failure
122	31	Failure	Failure
110	12	Failure	Failure
304	11	Failure	Success
283	15	Failure	Success

Confusion Matrix

We can express the **predicted** and **actual** outcomes in a two-by-two table instead.

	Success	Failure
Predicted Success	1	2
Predicted Failure	1	3

Accuracy

Accuracy is given

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{All Predictions}}$$

Accuracy

Accuracy is given

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{All Predictions}}$$

Example:

	Success	Failure
Predicted Success	1	2
Predicted Failure	1	3

Accuracy

Accuracy is given

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{All Predictions}}$$

Example:

	Success	Failure
Predicted Success	1	2
Predicted Failure	1	3

$$\text{Accuracy} = \frac{1+3}{1+2+1+3} = \frac{4}{7}$$

Accuracy Can Be Misleading!

Generally, **accuracy** is actually **NOT** a good measure.

Accuracy Can Be Misleading!

Generally, **accuracy** is actually **NOT** a good measure.

Example: Suppose you want to detect credit card fraud to decide which accounts to shut down. It is estimated that 99.9% of accounts are not fraudulent.

Accuracy Can Be Misleading!

Generally, **accuracy** is actually **NOT** a good measure.

Example: Suppose you want to detect credit card fraud to decide which accounts to shut down. It is estimated that 99.9% of accounts are not fraudulent.

Predicting every single account to be not fraudulent gives you a 99.9% accuracy ... but this isn't helpful at all.

Precision

Precision is given by

$$\text{Precision} = \frac{\text{True Positive}}{\text{All Positive Predictions}}.$$

In words, **precision** is how correct your positive predictions are.

Precision

Precision is given by

$$\text{Precision} = \frac{\text{True Positive}}{\text{All Positive Predictions}}.$$

In words, **precision** is how correct your positive predictions are.

Example:

	Success	Failure
Predicted Success	1	2
Predicted Failure	1	3

Precision

Precision is given by

$$\text{Precision} = \frac{\text{True Positive}}{\text{All Positive Predictions}}.$$

In words, **precision** is how correct your positive predictions are.

Example:

	Success	Failure
Predicted Success	1	2
Predicted Failure	1	3

$$\text{Precision} = \frac{1}{1+2} = 0.333$$

Recall

Recall is given by

$$\text{Recall} = \frac{\text{True Positive}}{\text{All Actual Positives}}.$$

In words, **recall** is how many **positives** you were able to recover.

Recall

Recall is given by

$$\text{Recall} = \frac{\text{True Positive}}{\text{All Actual Positives}}.$$

In words, **recall** is how many **positives** you were able to recover.

Example:

	Success	Failure
Predicted Success	1	1
Predicted Failure	1	3

Recall

Recall is given by

$$\text{Recall} = \frac{\text{True Positive}}{\text{All Actual Positives}}.$$

In words, **recall** is how many **positives** you were able to recover.

Example:

	Success	Failure
Predicted Success	1	1
Predicted Failure	1	3

$$\text{Recall} = \frac{1}{1+1} = 0.5$$

What's Good?

In general, there is no standard for what constitutes a "good" **precision** or **recall**. It all depends on your frame of reference.

Example: Suppose you want to detect credit card fraud to decide which accounts to shut down. It is estimated that 99.9% of accounts are not fraudulent.

What's Good?

In general, there is no standard for what constitutes a "good" **precision** or **recall**. It all depends on your frame of reference.

Example: Suppose you want to detect credit card fraud to decide which accounts to shut down. It is estimated that 99.9% of accounts are not fraudulent.

If we chose fraudulent accounts **at random**, we'd have a **precision** of about 0.1%. So, a **model with 10% precision** would be excellent, because that means the model gives you a **100x lift** over random chance.

What's Good?

Example: Suppose you want to detect credit card fraud to decide which accounts to shut down. It is estimated that 99.9% of accounts are not fraudulent. Would a model with **10% recall** be considered good?

What's Good?

Example: Suppose you want to detect credit card fraud to decide which accounts to shut down. It is estimated that 99.9% of accounts are not fraudulent. Would a model with **10% recall** be considered good?

If we chose 1% of all accounts **randomly** to flag as fradulent, then our **recall** would be 1%. If our model flags 1% of all counts as fraudulent and our recall is 10%, it is giving us a **10x lift** over random chance.

What to Use?

Which metric matters more depends on what you care about.

Recall: Suppose a car company wants to decide which cars to buy ad time for in the next Super Bowl. They have a dataset of cars promoted in past Super Bowl ads, and information about the **miles per gallon** (`mpg`) and **horsepower** (`hp`) of each car. In addition, they know which ad campaigns were deemed **"Successful"** or **"Unsuccessful."** Which cars should the company buy ad time for?

What to Use?

Which metric matters more depends on what you care about.

Recall: Suppose a car company wants to decide which cars to buy ad time for in the next Super Bowl. They have a dataset of cars promoted in past Super Bowl ads, and information about the **miles per gallon** (`mpg`) and **horsepower** (`hp`) of each car. In addition, they know which ad campaigns were deemed **"Successful"** or **"Unsuccessful."** Which cars should the company buy ad time for?

In this case, we might want to use **precision** because Super Bowl ads are expensive.