Lecture 6: Intro to Classifiers

INFO 1998: Introduction to Machine Learning

CDS Education

Apply to Cornell Data Science! 📣

- All subteams are recruiting freshmen this semester!
  - Deadline: **October 17th, 11:59pm**
  - Don’t forget to also submit the College of Engineering application.

- Application Link: https://cornelldata.science/recruitment

- If you’re enjoying this class…
  - you’ll LOVE being on CDS 👀

Subteam UTea trip!

Agenda

1. What is a Classifier?
2. K-Nearest Neighbors Classifier
3. Review of Underfitting v. Overfitting
4. Confusion Matrices

What are Classifiers?

What are Classifiers?

Classifiers are able to help answer questions like…

- "What species is this?"
- "What major is a student in based on their classes?"
- "Which Hogwarts House do I belong to?"
- "Am I going to pass this class?"

What are Classifiers?

- Classifiers predict the class/category of a set of data points. This class/category is based off of the target variable we are looking at.

- Difference between linear regression and classifiers
    - Linear regression is used to predict the value of a **continuous variable**
    - Classifiers are used to predict **categorical or binary variables**

K-Nearest Neighbors Classifier

What is the KNN Classifier?

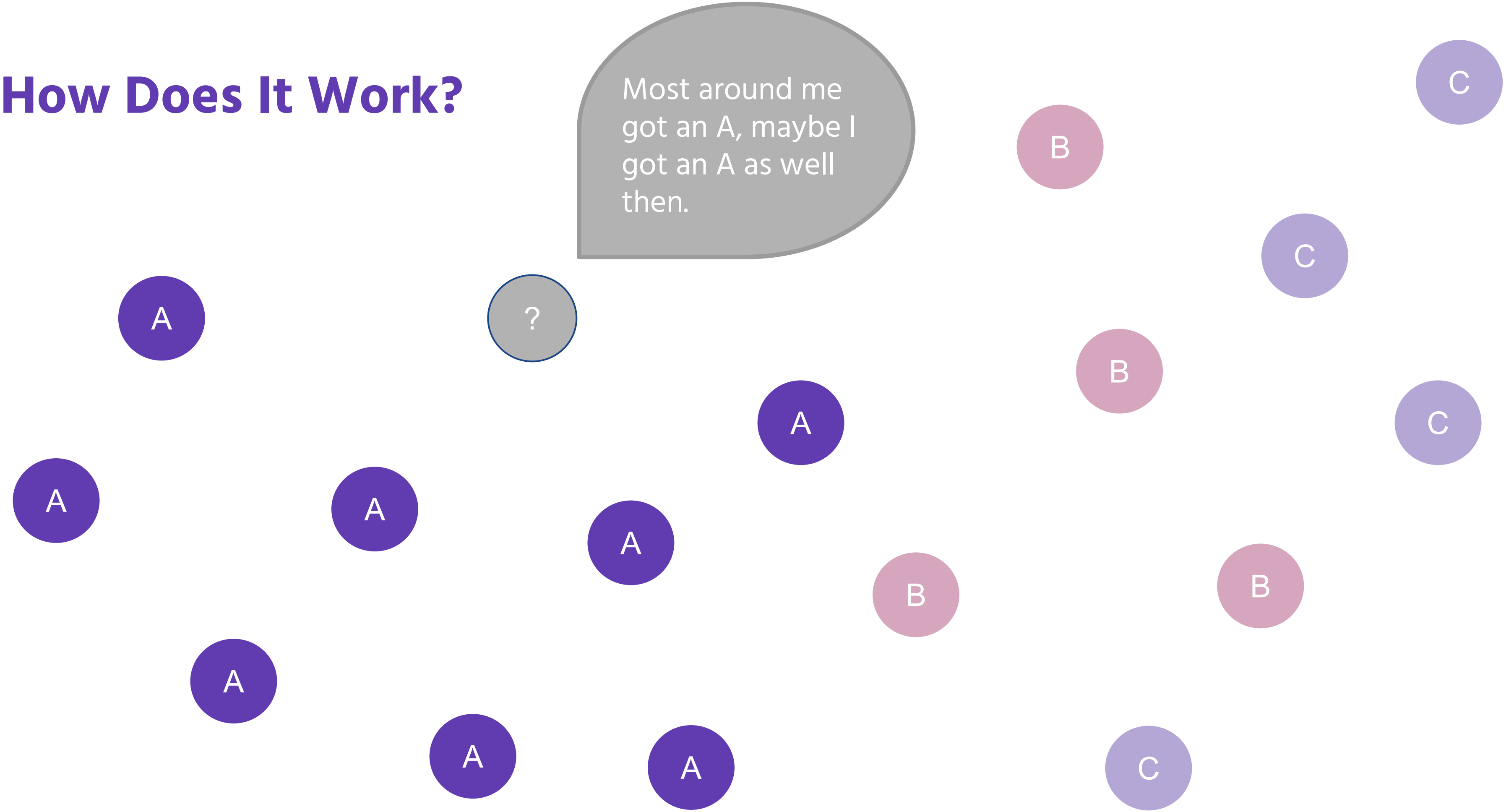
- Lazy learner classifier
- Easy to interpret
- Fast to calculate
- Good for coarse analysis

How Does It Work?

Uses the k (a user specified value) nearest data points to predict the unknown one

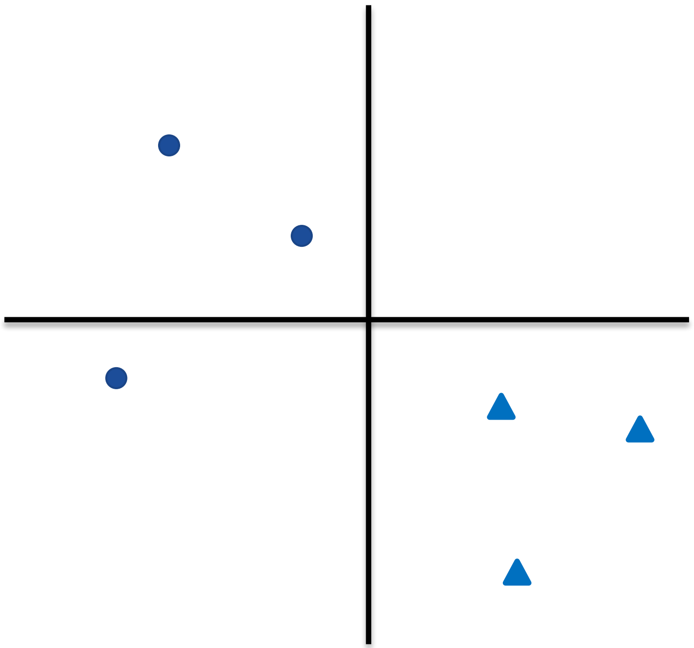
- A simple assumption: the values **nearest** to a data point are **similar** to it
- k is a **hyperparameter** of the KNN model
  - a parameter which affects the training process

How Does It Work?



How Does It Work? (Step-By-Step Example)

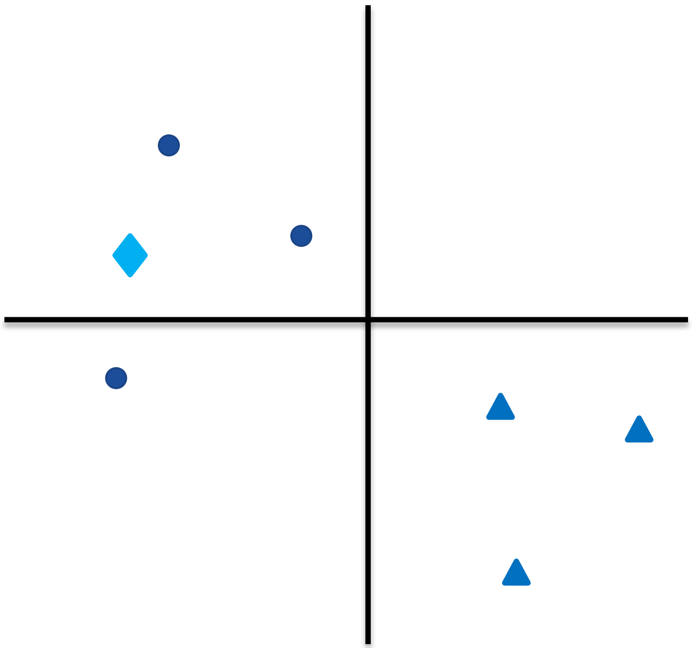
Define a *k* value (in this case k = 3)



How Does It Work? (Step-By-Step Example)

Define a *k* value (in this case k = 3)

Pick a point to predict (blue diamond)

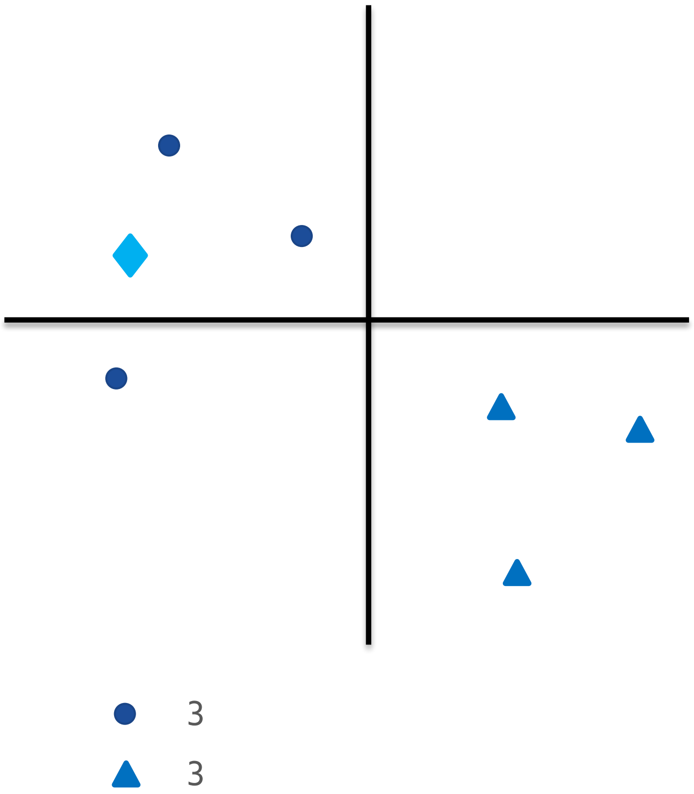


How Does It Work? (Step-By-Step Example)

Define a *k* value (in this case k = 3)

Pick a point to predict (blue diamond)

Count the number of closest points



How Does It Work? (Step-By-Step Example)

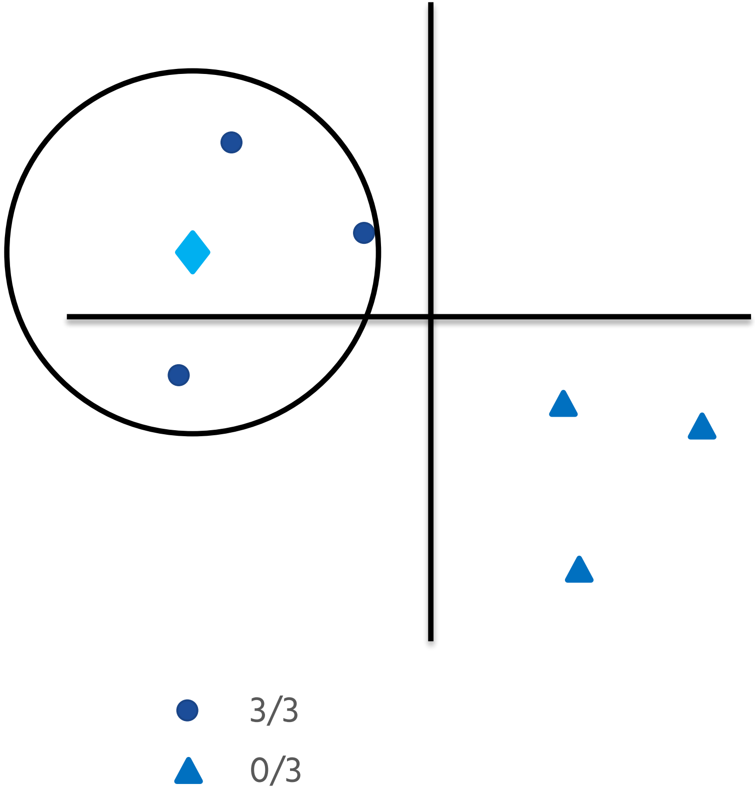
Define a *k* value (in this case k = 3)

Pick a point to predict

(blue diamond)

Count the number of closest points

Increase the radius until the number of points in circle adds up to 3



How Does It Work? (Step-By-Step Example)

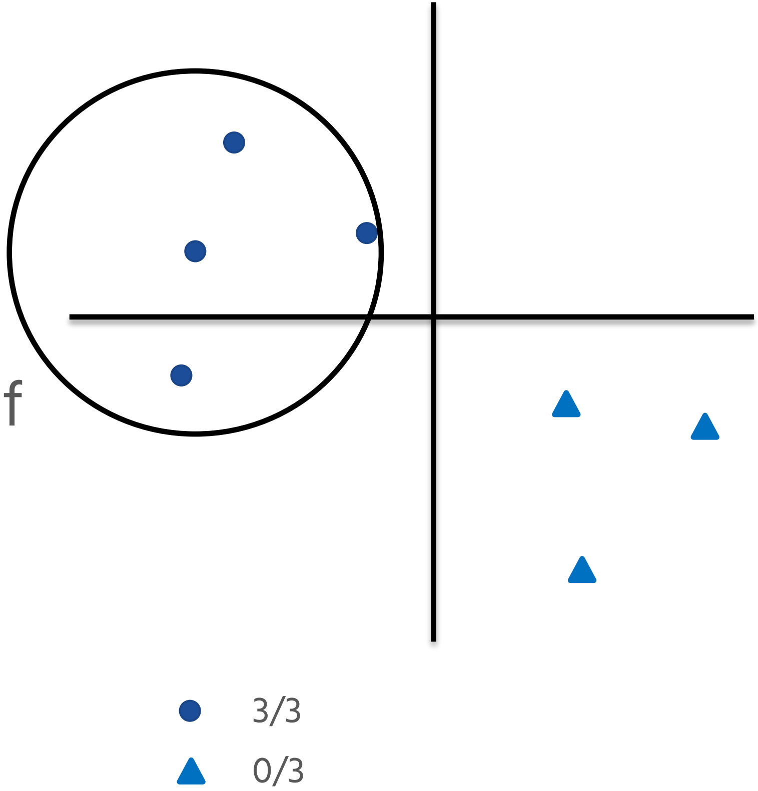
Define a *k* value (in this case k = 3)

Pick a point to predict (blue diamond)

Count the number of closest points

Increase the radius until the number of points within the radius adds up to 3

Predict the blue diamond to be a blue circle!



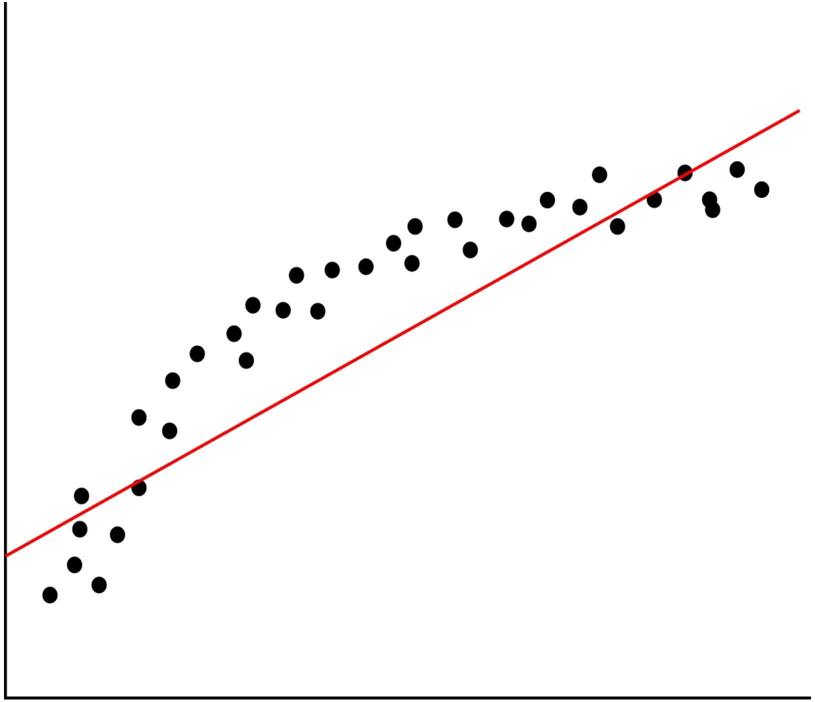
Demo

Underfitting v. Overfitting

Underfitting

Underfitting means we have high bias and low variance.

- Lack of relevant variables/factor
- Imposing limiting assumptions
  - Linearity
  - Assumptions on distribution
  - Wrong values for parameters

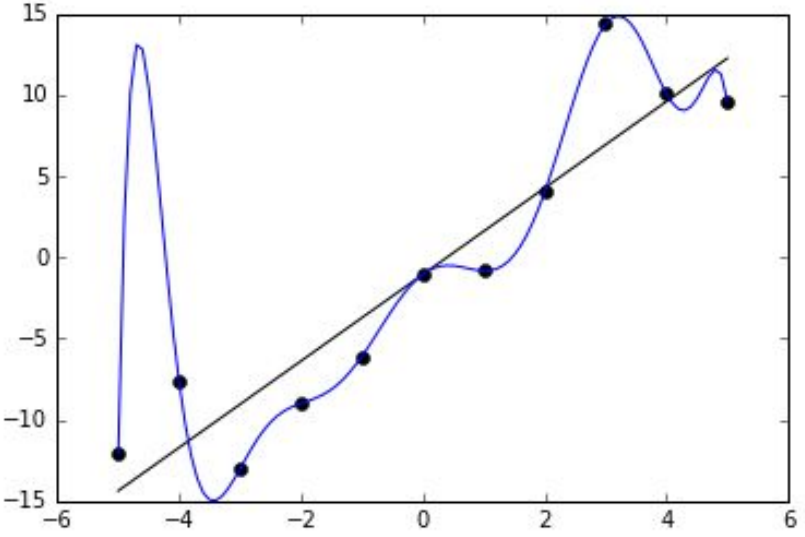


Source

Overfitting

Overfitting means we have low bias and high variance.

- Model fits too well to specific cases
- Model is over-sensitive to sample-specific noise
- Model introduces too many variables/complexities than needed

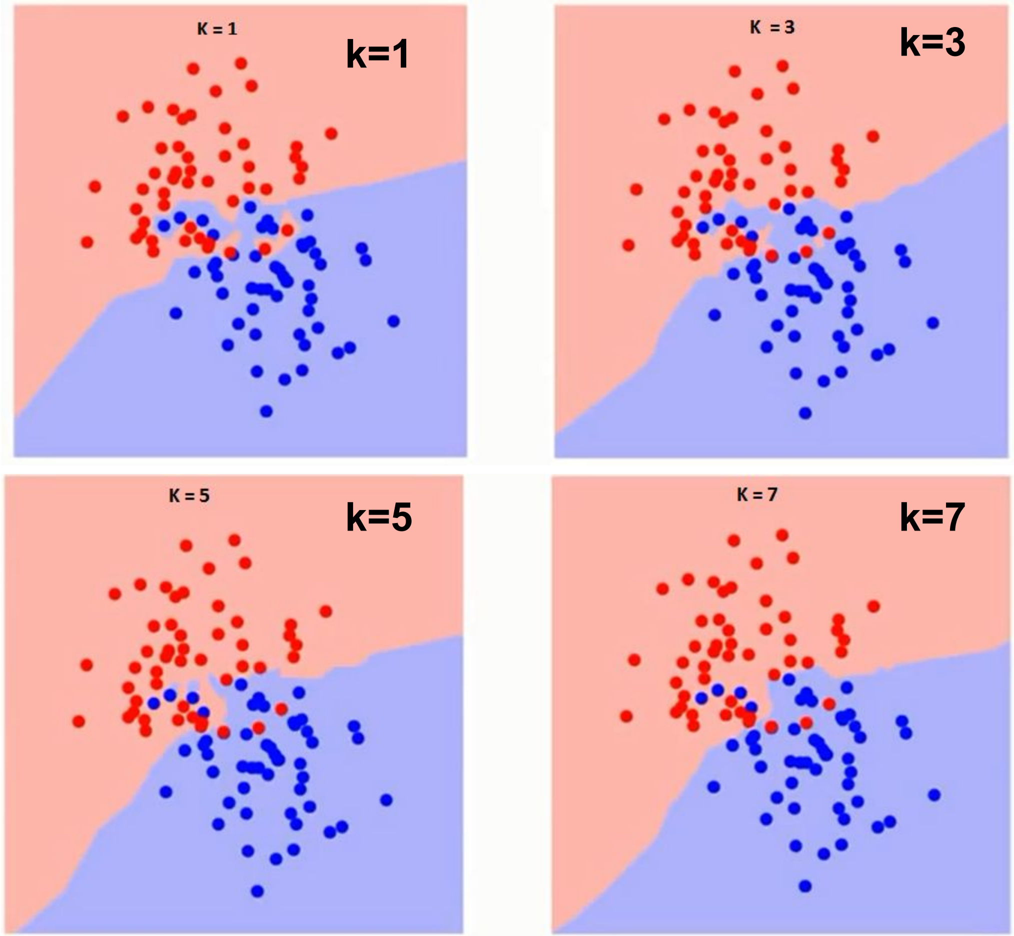


Source

Relationship Between k and Fit

The **k** value you use has a relationship to the fit of the model

A higher k gives a smoother line, but too large of a k and it is the average of all the data (or the label that is most common/likely)



Confusion Matrix

What is a Confusion Matrix?

Table used to describe the performance of a classifier on a set of binary test data for which the true values are known

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Sensitivity

Called the **true positive rate**

Tells us how many positives are correctly identified as positives

Optimize for: Initial diagnosis of fatal disease

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Sensitivity = True Positive/ (True Positive + False Negative)

Specificity

Called the **true negative rate**

Tells us how many negatives are correctly identified as negatives

Optimize for: testing for a disease with a risky treatment

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Specificity = True Negative/ (True Negative + False Positive)

Question

Which is an example of when you would want **higher specificity?**

A. DNA tests for a death penalty case
B. Deciding which iPhone to buy
C. Airport security

Overall Accuracy

Proportion of correct predictions

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Accuracy = (True Positive + True Negative) / Total

Overall Error Rate

Proportion of incorrect predictions

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Error = (False Positive + False Negative) / Total

Precision

Proportion of correct positive predictions among all positive predictions

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Precision = True Positive / (True Positive + False Positive)

Coming Up

- **Assignment 5:** Due **Friday 10/18** at 11:59pm!
- **Assignment 6**: Due next Wednesday 10/23
- **Mid-Semester Check-In:** Details on ED! Complete by Wednesday 10/23.
- **Next Lecture**: Supervised Learning Pt. 1