Machine Learning Using Python

Xiyuan Ge
University of Washington

Foster School of Business
UNIVERSITY of WASHINGTON

Self Intro

Xiyuan ('C1') Ge

- UNC-Chapel Hill, NYU, UVa PhC in Machine Learning
- Work experience in finance
- Research
  - Online review and its implication on supply chain
  - Customer OOS substitution in service industry
  - Applied ML: RNA sequencing & structure prediction

Outline

- Objectives
- Linear Regression
- K-Nearest Neighbor (KNN)

Objectives

- Review basic machine learning algorithms
- Familiarize with Google Colab environment for Python programming
- Help transition to advanced topics

Files for the Workshop

https://colab.research.google.com/drive/1aHoNqPOZAMXsglSqfo3bD9tg4KxgVIKp

- Save a copy to your own local/Google drive

Outline

- Objectives
- Linear Regression
- K-Nearest Neighbor (KNN)

Linear Regression Example

- Data set: ToyotaCorolla.csv
- Examine the factors affecting the price

Price	Age	KM	FuelType	HP	Metallic	Automatic	cc	Doors	Weight
13500	23	46986	Diesel	90	1	0	2000	3	1173
13750	23	72937	Diesel	90	1	0	2000	3	1198
13950	24	41711	Diesel	90	1	0	2000	3	1219
14950	26	48000	Diesel	90	0	0	2000	3	1321

Linear Regression

- Run a linear model using Python
- What is your linear model?

Interpret the Model

- Coefficient
  - Estimated marginal effect of input variable on output variable (price)
- Standard Error
  - It measures the estimation precision of the coefficient
- p-value
  - It measures how likely the coefficient has no effect on the outcome
  - When p-value is too high, e.g. more than 0.10 or 0.05, the effect of a variable becomes insignificant
- R-squared
  - A measure of the fit of the model. Proportion of total variation in the outcome variable explained by the model

Explanatory Modeling vs. Predictive Modeling

- Explanatory
  - Explain relationship between explanatory (independent) variables and dependent variable
  - Fit the data well and understand the contribution of explanatory variables to the model
  - Performance measures
    - R^2, residual analysis, p-values
- Predictive
  - Predict target values in other data where we have predictor values, but not target values
  - Classic data mining context and model goal is to optimize predictive accuracy
  - Performance measure is assessed on validation data
  - Explaining role of predictors is not primary purpose (but useful)

Outline

- Objectives
- Linear Regression
- K-Nearest Neighbor (KNN)

Predictive Modeling

- Supervised learning
  - Computer vision, speech recognition, self-driving cars, AlphaGo
  - Goal: Predict a single “target” or “outcome” variable
  - Training data, where target value is known
  - Methods: Classification and Prediction
  - Using the trained model to score new observations, where value is not known
- Unsupervised learning
  - Goal: Segment data into meaningful segments; detect patterns
  - There is no target (outcome) variable to predict or classify
  - Methods: Association rules, data reduction & exploration, visualization

Supervised: Classification

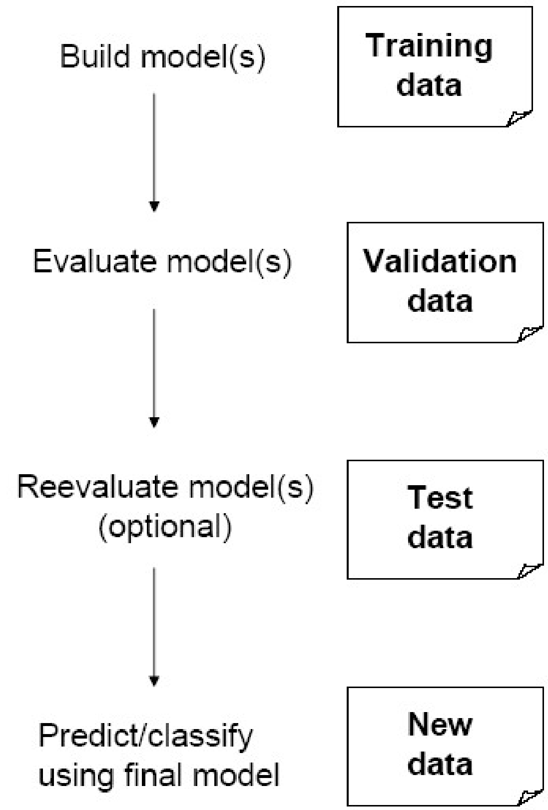
- Goal: Predict categorical target (outcome) variable
- Examples: Purchase/no purchase, fraud/no fraud, creditworthy/not creditworthy…
- Each row is a case (customer, tax return, applicant)
- Each column is a variable
- Target variable is often binary (yes/no)

Supervised: Prediction

- Goal: Predict numerical target (outcome) variable
- Examples: sales, revenue, performance
- Each row is a case (customer, tax return, applicant)
- Each column is a variable
- Taken together, classification and prediction constitute "predictive analytics"

Partitioning the Data

- Problem: How well will our model perform with new data?
- Solution: Separate data into two parts
  - Training partition to develop the model
  - Validation partition to implement the model and evaluate its performance on "new" data
- Addresses the issue of overfitting



Sampling

- The function np.random.seed()
- The purpose is to have reproducible results so that we can debug our program

np.random.seed() will make the results reproducible

Measuring the Prediction Model

- Root mean square error (RMSE)

$$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Characteristics of KNN

- Data-driven!
- Not model-driven
- A “purer” machine learning algorithm

Basic Idea

- For a given record to be classified or predicted, identify nearby records
- "Near" means records with similar predictor values X_1, X_2, … X_p
- Classify the record as whatever the predominant class is among the nearby records (the "neighbors")

How to measure nearby?

The most popular distance measure is **Euclidean distance**

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + \left(x_p - y_p\right)^2}$$

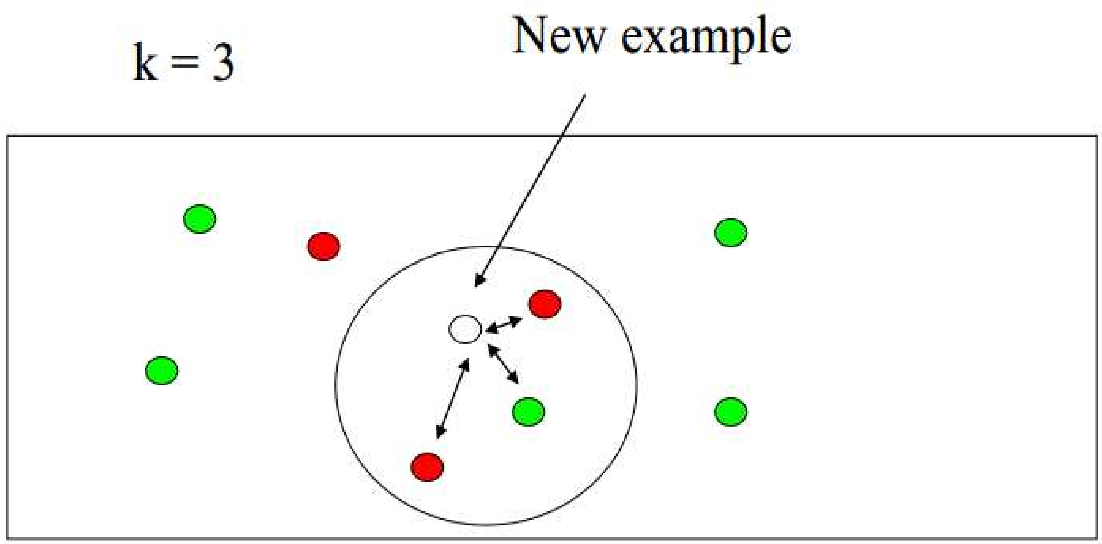
W Foster School of Business

What does k mean?

- *K* is the number of nearby neighbors to be used to classify the new record
  - *K*=1 means use the single nearest record
  - *K*=5 means use the 5 nearest records

Example

- Find the k-nearest neighbors and have them vote. Here, k=3.
- By taking more than one neighbors, the impact of outliers can be reduced.
- A practical note: It is typical to use odd number for k to avoid ties



W Foster School of Business

How to choose k?

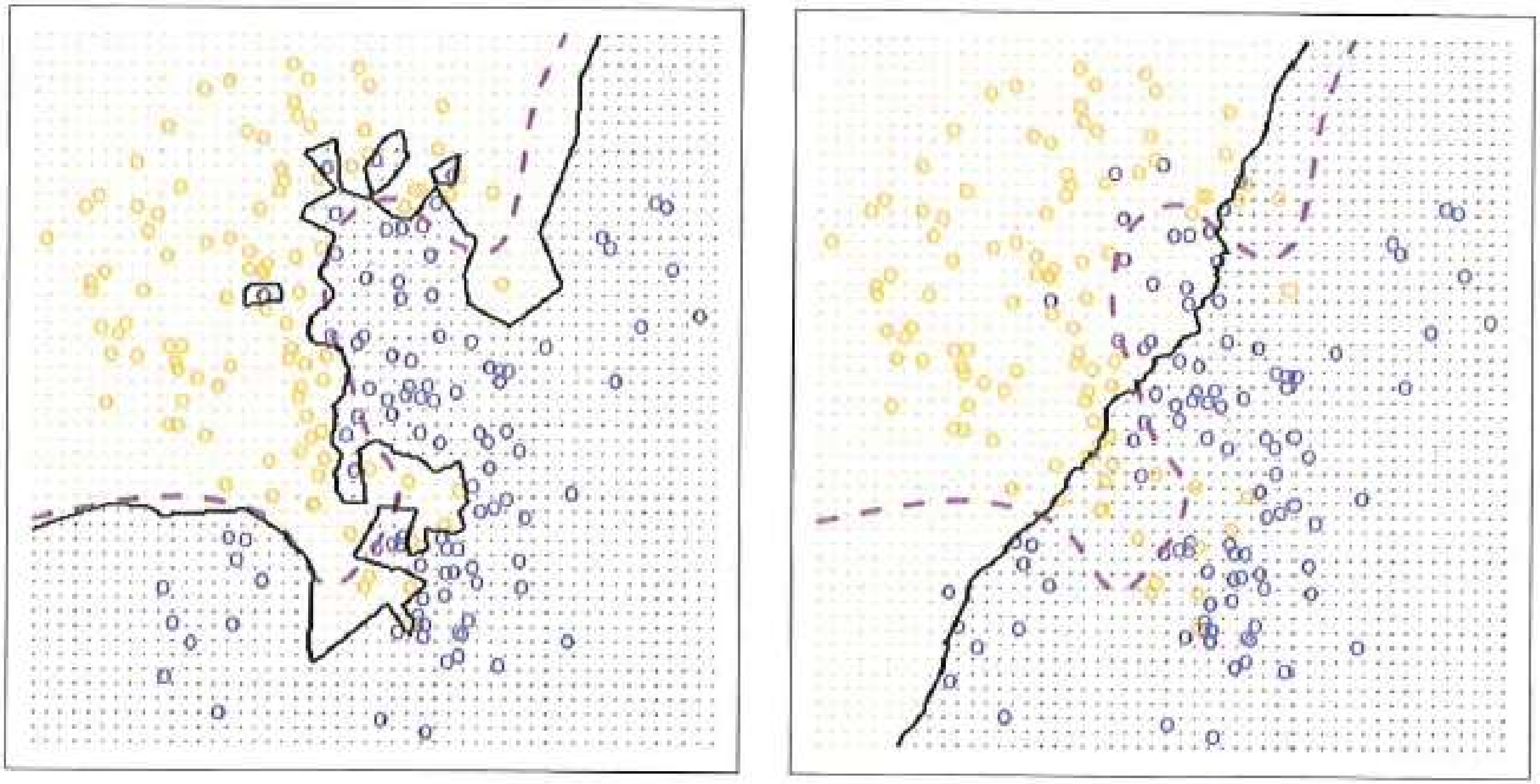
- Typically choose the value of *k,* which has the lowest error rate in validation data

Low *k* vs. High *k*

- Low values of *k* (1, 3, …) capture local structure in data (but also noise)
- High values of *k* provide more smoothing, less noise, but may miss local structure

KNN: K=1

KNN: K=100



Using K-NN for Classification

- Instead of average of response values, use majority vote among neighbors

Measuring Classification: Confusion Matrix and Error Rate

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	126	68
0	26	1780

Error Report			
Class	# Cases	# Errors	% Error
1	194	68	35.05
0	1806	26	1.44
Overall	2000	94	4.70

W | Foster School of Business

Homework (optional)

- Implement K-NN classification model using a dataset from last quarter