Week 2

Agenda

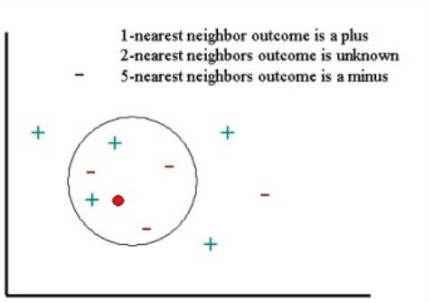
1. Async Review: KNN
2. Case study: house price prediction
3. Parameters & Hyperparameters
4. Terminology and data sets
5. Breakout: Project 1
6. If time notebooks

Warm up

1. **Performance on test data is always worse than performance on training data, true or false?**
2. **Performance on test data is always more meaningful than performance on training data, true or false?**
3. **Ideally you should use your test data**
   - exactly once.
   - only for error analysis.
   - for repeat experiments.
4. **Simple decision boundaries are more likely to generalize to new data, true or false?**
5. **Find L1 distance for x1 = [1,2,2], x2 = [4,4,2]**
6. **MNIST: number of possible inputs is:**
   28 x 28
   28 x 28 x 10
   256^(28x28)

KNN

KNN Review



Nearest Neighbor Example: Results

- Test set size = 10,000 digits

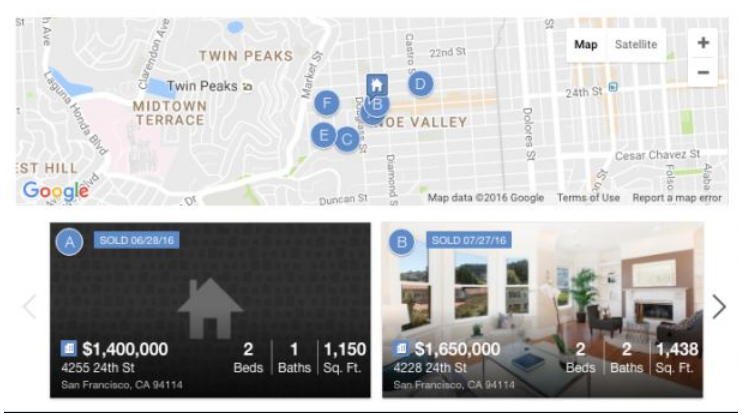
k = 1; Euclidean (L2) distance

Training	Error %	Time (secs)
100	30.0	0.38
1,000	12.1	2.34
10,000	5.3	28.7
60,000	2.7	2202
Deskewing	2.3	
Blurring	1.8	
Pixel shifting	1.2	

- State-of-the-art error rate: 0.3%

1. Describe the NN Algorithm?
2. What changes with K-NN?
3. How much memory is used to store the trained model?
4. How computationally fast are predictions?
5. In general, what is the difference between regression and classification problems?
6. What is edited KNN?

Natural Domains for KNN

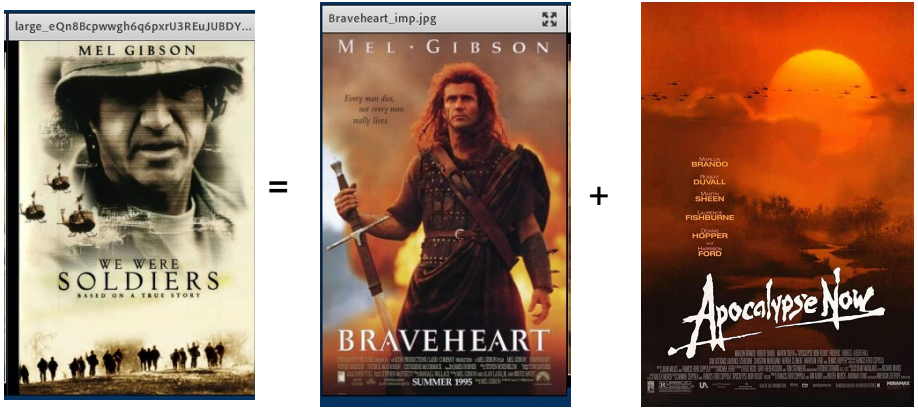


Comparables

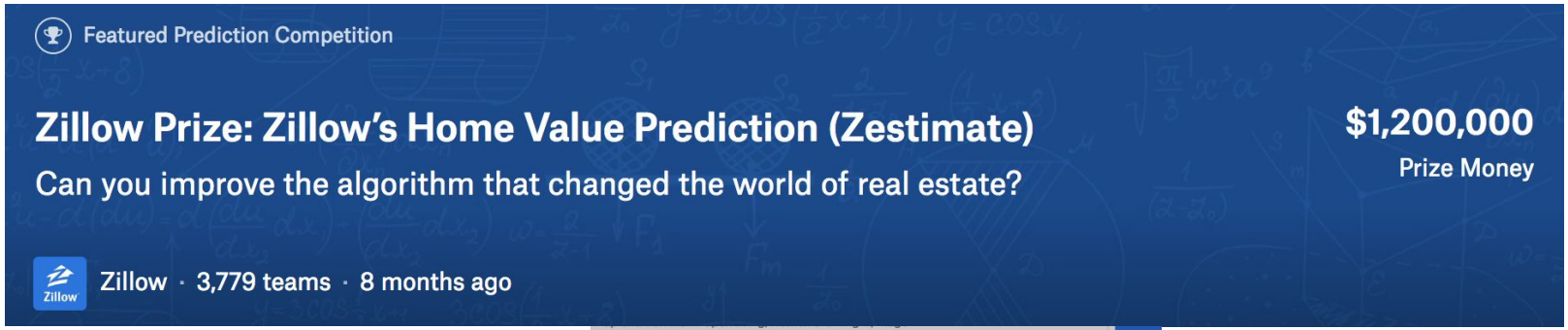
From Wikipedia, the free encyclopedia

Comparables (or **comps**) is a real estate appraisal term referring to properties with characteristics that are similar to a subject property whose value is being sought. This can be accomplished either by a real estate agent who attempts to establish the value of a potential client's home or property through market analysis or, by a licensed or certified appraiser or surveyor using more defined methods, when performing a real estate appraisal.

1. When have you seen people naturally use NN-like algorithms?
2. From a data perspective, when do you think you might want to use NN?



House prices case study: what goes into that?



"Zestimates" are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And, by continually improving the median margin of error (from 14% at the onset to 5% today), Zillow has since become established as one of the largest, most trusted marketplaces for real estate information in the U.S. and a leading example of impactful machine learning.

fireplace & Sound views. Spacious kitchen w/ slab granite surfaces & center island. Huge master suite with Jacuzzi tub & separate shower. Features; hdwd floors, all

Parameters and Hyperparameters

Parameters

$$price = \alpha + \beta(bedrooms) + \gamma(bathrooms)$$

1. What are the parameters in this model?
2. What is the role of data in parametric modeling?
3. What might make one set of parameter values better than another?

Hyperparameters

Price = KNN(house, trainingData, similarity, K)

1. What are the parameters in this model?
2. What are hyperparameters?
3. What happens when K == size of training Data?
4. Is there a Hyperparameter in the last example?

Similarity Measures

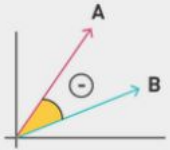
Similarity/Distance Measures

Distance Metrics

$$L^n(x_1, x_2) = \sqrt[n]{\sum_{i=1}^{\#\dim} |x_{l,i} - x_{2,i}|^n}$$

- For numeric features:
  - Manhattan distance
  - Euclidean distance
  - L^n-norm

$$\frac{A \cdot B}{\|A\| \ \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$



- For cosine similarity:
  - Only the angle between the vectors matters.

Edit Distance: Example

TGCATAT → ATCCGAT in 5 steps

TGCATAT → (delete last T)
TGCATA → (delete last A)
TGCAT → (insert A at front)
ATGCAT → (substitute C for 3rd G)
ATCCAT → (insert G before last A)
ATCCGAT (Done)

What is the edit distance? 5?

1. What are some similarity measures you know?
2. What makes for a 'good' similarity measure?

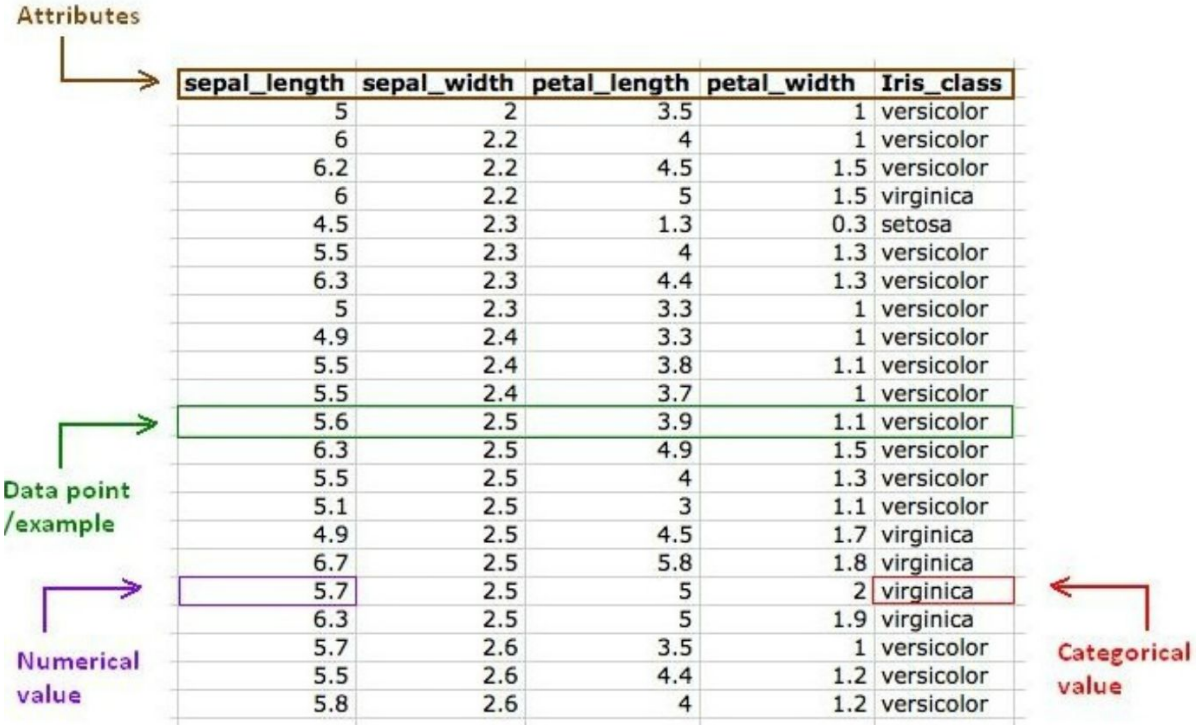
Measures and Metrics

1. $d(x, y) \geq 0$ (*non-negativity*, or separation axiom)
2. $d(x, y) = 0$ if and only if $x = y$ (coincidence axiom)
3. $d(x, y) = d(y, x)$ (*symmetry*)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (*subadditivity / triangle inequality*).

1. What is a 'metric'? Semi-metric?
2. What is an example of a metric?

Machine Learning Workflows/Terminology

Design Matrix



sepal_length	sepal_width	petal_length	petal_width	Iris_class
5	2	3.5	1	versicolor
6	2.2	4	1	versicolor
6.2	2.2	4.5	1.5	versicolor
6	2.2	5	1.5	virginica
4.5	2.3	1.3	0.3	setosa
5.5	2.3	4	1.3	versicolor
6.3	2.3	4.4	1.3	versicolor
5	2.3	3.3	1	versicolor
4.9	2.4	3.3	1	versicolor
5.5	2.4	3.8	1.1	versicolor
5.5	2.4	3.7	1	versicolor
5.6	2.5	3.9	1.1	versicolor
6.3	2.5	4.9	1.5	versicolor
5.5	2.5	4	1.3	versicolor
5.1	2.5	3	1.1	versicolor
4.9	2.5	4.5	1.7	virginica
6.7	2.5	5.8	1.8	virginica
5.7	2.5	5	2	virginica
6.3	2.5	5	1.9	virginica
5.7	2.6	3.5	1	versicolor
5.5	2.6	4.4	1.2	versicolor
5.8	2.6	4	1.2	versicolor

Building Datasets

Digit Classification

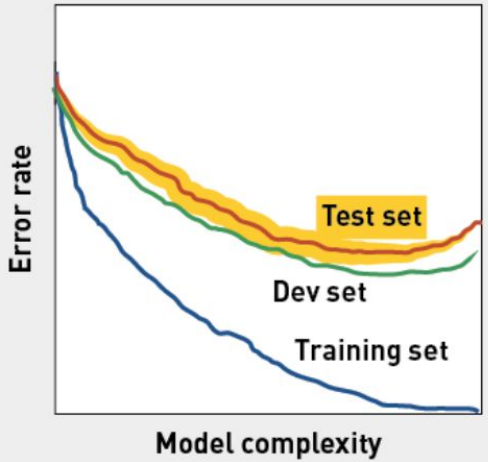
- MNIST digit data set
  - Widely used test data for classification systems.
  - Contains 70,000 labeled digits.
    - 60,000 for training
    - 10,000 for testing
  - Half are from Census Bureau workers.
  - Half are from high school students.
  - Data should be randomized to include samples from the workers and students.
  - Data were scaled and centered.

1. What makes MNIST a good/bad benchmark dataset? Could you use it to "learn" what a cat looks like?
2. How would you create a benchmark dataset for home price prediction?

Workflow

Development Data

- Errors in the test data should be examined.
  - However, examining the test set introduces bias.



1. What is the role of a dev dataset?
2. What is the role of error analysis?

Error Analysis

Cross-Validation

- Split a single set of data into training and test sets in many different ways.
- Important when there are a small amount of data.
- We want to use as much data as possible for training and as much data as possible for testing.
- **Jack-knife:** Split data into training and test data.
- **Leave-one-out:** Use all of the data for training except for one.
  - Repeated until each data point has been left out
- **Randomized splits:** jack-knife split with random partitions.

1. When might you consider using cross-validation?
2. How do you deploy a model Developed using CV?

Notebook from git and Scikit learn

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html