SI 670 Notes

Suggested books

- Introduction to Machine Learning with PythonA Guide for Data ScientistsBy Andreas C. Müller and Sarah Guido
- Deep Learning with Pythonby Francois Chollet

Top libraries

- Scikit-learn
- SciPy
- Numpy
- Pandas
- Matplotlib

Cycle

- Feature representation
- Training
- Evaluation
- Refine cycle (hyperparameterization)

Data quality checks

- Min/max summaries
- Wrong data type, units
- Equal class reppresentation
- Outliers
- Data distribution
- Correlations among variables

KNN Notes

Category

- Supervised
  - Classification
  - Regression

High level algorithm

Given a training set X-train with labels y-train, and given a new instance x-test

1. Find the observations that resamble x-test that are in X-train. Call this set of observation(s) Xnn
2. Get the labels of Ynn for the instances in Xnn
3. Predict label for x-test by combining the labels Ynn (majority vote).

Parameters

- Distance metric (Euclidian)
- Choice of k (k=1 very flexibel, k=100 rigid)
- Weighting function (neighbors that are far less influence on final prediction)

Evaluation

- Accuracy (correctly predicted / total observations) (for classification)
- R^2 (for regression, measure how does the data fit the model 0-1)

Extras

- Ensure that all observations are on the same scale
  - if not standardize them (standard scalar)

Classification

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X_C1, y_C1, random_state = 0)
knnc = KNeighborsClassifier(   n_neighbors = 5).fit(X_train, y_train)
print(knnc.predict(X_test))
print('Accuracy test score: {:.3f}'.format(knnc.score(X_test, y_test))
```

Regression

```
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test =train_test_split(X_R1, y_R1, random_state = 0)
knnreg  = KNeighborsRegressor(n_neighbors = 5).fit(X_train, y_train)
print(knnreg.predict(X_test))
print('R-squared test score: {:.3f}'.format(knnreg.score(X_test, y_test)```
```