Preprocessing data

Feature Scaling

Daniel Shin

What is feature scaling?

Feature scaling is a method used to **standardize** the **range** of independent variables or features of data. In data processing, it is also known as data **normalization** and is generally performed during the data **preprocessing** step.

Why use feature scaling?

- Many estimators require it
  - Regression
  - SVM
  - kNN
  - Neural networks
- Image processing

How to use feature scaling?

- Methods
  - Calculate z-scores of your feature variables

$$z = \frac{x - \mu}{\sigma}$$

  - Min-max method

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

When to use feature scaling?

- Andrew Ng
  - Okay if variables in the [-⅓ , ⅓] to [-3 , 3] range
  - Otherwise, scale

sklearn.preprocessing

- *preprocessing.scale*
  - by default scales feature to mean 0 and variance 1
  - i.e calculates the Z-score
- *preprocessing.normalize*
  - scales each feature to unit vector
  - useful for text classification/clustering

Results

Pre-scaled	accuracy	precision	recall	f1
LogisticReg	0.838	0.857	0.788	0.817
SVMC	0.539	0.000	0.000	0.000
GaussNB	0.828	0.828	0.802	0.811
DecisionTree	0.703	0.688	0.736	0.690
RandomForest	0.801	0.840	0.685	0.741
kNN9	0.653	0.637	0.576	0.605

Scaled	accuracy	precision	recall	f1
LogisticReg	0.835	0.850	0.787	0.813
SVMC	0.828	0.846	0.773	0.803
GaussNB	0.828	0.828	0.802	0.811
DecisionTree	0.714	0.720	0.729	0.693
RandomForest	0.818	0.845	0.700	0.773
kNN9	0.825	0.841	0.766	0.800

Actually this is wrong

Fit scaler only on training set

```
std_scale = preprocessing.StandardScaler().fit(X_train)
X_train_std = std_scale.transform(X_train)
X_test_std = std_scale.transform(X_test)
```

also MinMaxScaler()