16. Machine Learning Introduction

3ikakke

Friday 22nd April 2022

Outline

- Objectives
- Machine Learning
- Supervised Learning
- Unsupervised Learning
- Other Machine Learning Algorithms
- Approach to making predictions
- Using SKLearn
- Refresher on probability
- Evaluating Classification Problems
- Evaluating Regression Problems

Objectives

- Understand broadly the concept of machine learning
- Understand the difference between supervised and unsupervised learning
- Understand data pre-processing

Machine Learning

- Basically using mathematical approaches to make predictions
- These predictions may be based on some prior knowledge captured in the data
- Where we use labels within the data to determine what unlabeled data points are we call that supervised learning
- Where there are no labels but we want to determine differences in the data we call that unsupervised learning
- Machine learning is the basis for deep learning

Supervised Learning

- These are based on labels that exist
- The variables we want to predict are called labels while those used in predicting are called features
  - Labels vs Features
  - Dependent variables vs Independent variables
  - Outcome vs Exposure
- When the label is categorical we use classification algorithms
- When the label is numeric we use regression algorithms

Unsupervised Learning

- These are not based on labels
- The algorithms attempt to find similarities and differences in the data
- Examples
  - Clustering algorithms
  - Dimensionality reduction algorithms (Principal Component Analysis and Linear Discriminant Analysis)

Other Machine Learning Algorithms

- Ensemble algorithms
  - Random Forest Regression (bagging algorithm)
  - Random Forest Classification (bagging Algorithm)
  - XGBoost (boosting algorithm)
- Recommender systems
  - Facebook friend suggestion
  - Netflix movie recommendation
- More!

Approach to making predictions

- Usually you need to have thee parts of the data
  - Training set
  - Validation set
  - Test set
- The training-validation-testing lifecycle
- Be careful of overfitting

Using SKLearn

- The parent module for machine learning

- Lets refer to the features as x and the labels as y

- Preprocessing
  1. Import module:
     ```
     from sklearn.cross_validation import train_test_split
     ```
  2. Create the split:
     ```
     xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size=0.75)
     ```
- Using the algorithm
  1. Import the required model from a family
     ```
     from sklearn.family import Model
     ```
  2. Instantiate the model
     ```
     model = Model()
     ```
  3. Fit the model:
     ```
     model.fit(xtrain, ytrain)
     ```
  4. Predict:
     ```
     model.predict(xtest)
     ```

Refresher on probability

	Actual True	Actual False
Predited True	a	b
Predicted False	c	d

- $P(PredictedTrue \cap ActualTrue) = \mathbf{a} = True\ Positive$

- $P(PredictedTrue \cap ActualFalse) = \mathbf{b} = False\ Positive$

- $P(PredictedFalse \cap ActualTrue) = \mathbf{c} = False\ Negative$

- $P(PredictedFalse \cap ActualFalse) = \mathbf{d} = True\ Negative$

Evaluating Classification Problems

Table 2: Confusion Matrix

	Actual True	Actual False
Predited True	a	b
Predicted False	c	d

- Proportion of correct predictions = $\frac{a+d}{a+b+c+d} = Accuracy$

- $P(PredictedTrue|ActualTrue) = \frac{a}{a+c} = Recall$

- $P(ActualTrue|PredictedTrue) = \frac{a}{a+b} = Precision$

- Harmonic mean of *Recall* and *Precision* $= 2 * \frac{Recall*Precision}{Recall+Precision} = \textit{F1 Score}$

Evaluating Regression Problems

- Regression problems are those supervised learning problems in which the label is numeric (continuous)
- Mean Absolute Error (MAE) $= \frac{1}{n}\sum_{i}^{n} y_i - \hat{y}_i$
- Mean Square Error (MSE) $= \frac{1}{n}\sum_{i}^{n} (y_i - \hat{y}_i)^2$
- Root Mean Square Error (RMSE) $= \sqrt{\frac{1}{n}\sum_{i}^{n} (y_i - \hat{y}_i)^2}$

Review of objectives

- Understand broadly the concept of machine learning
- Understand the difference between supervised and unsupervised learning
- Understand data pre-processing

Q&A

There's no gist of the day!

Thanks for being part of the conversation!