Week 3

Friday, February 16, 2024 10:00 AM

K-Nearest Neighbors

Introduction to Classification

- Classification algorithms in ML
  - Decision Trees (ID3, C4.5, C5.0)
  - Naive Bayes
  - Linear Discriminant Analysis
  - *k*-Nearest Neighbor
  - Logistic Regression
  - Neural Networks
  - Support Vector Machines (SVM)

Basics

- What is KNN?
  - A method for classifying cases (or records) based on their closest neighbors.
  - Case or record is classified using majority vote of its neighbors.
  - Neighbors are taken from a set of cases for which we know the right classification.

- The KNN Algorithm
  1. Pick a value for K.
  2. Calculate the distance of unknown case from all cases.
  3. Select the K observations in the training data that are "nearest" to the unknown data point.
  4. Predict the response of the unknown data point using the most popular response value from the K nearest neighbors.
- KNN Can also be used for regression

Evaluation Metrics in Classification

- Jaccard index
  - Actual labels
  - Predicted labels

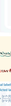
$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|}$$

$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$

- F1-score
  - Precision = TP / (TP + FP)
  - Recall = TP / (TP + FN)
  - F1-score is (up to now) calculated as the harmonic average of the precision and recall, calculated as:

$$F1\text{-score} = \frac{2 \times (precision \times recall)}{(precision + recall)}$$

- Log Loss
  - Performance of a classifier where the predicted output is a probability value between 0 and 1.



Decision Trees

Introduction to Decision Trees

- The small intuition behind a decision tree is to map out all possible decision paths in the form of a tree.
- Building a decision tree with not missing data

- Decision Tree Learning Algorithm
  1. Choose an attribute from your dataset.
  2. Calculate the significance of attribute in splitting of data.
  3. Split data based on the value of the best attribute.
  4. Go to step 1.

Building Decision Trees

- Which attribute is the best?

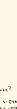
- Entropy
  - Measure of randomness or uncertainty.
  - Homo - homogeneous - adequate
  - The lower the Entropy, the less uniform the distribution, the purer the node.

- Which attribute is the best you can use?

Age	Income	Student	Credit	Buys
<30	High	No	Fair	No
<30	High	No	Excellent	No
31-40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31-40	Low	Yes	Excellent	Yes
<30	Medium	No	Fair	No
<30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
<30	Medium	Yes	Excellent	Yes
31-40	Medium	No	Excellent	Yes
31-40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

The tree with the higher Information Gain after splitting.

- What is information gain?
  - **Information gain** is the information that can increase the level of certainty after splitting.



- Entropy calculation
  - Number of characteristics = 2
  - Number of Buys = Yes = 9
  - Number of Buys = No = 5
  - Total number = 14

$$E(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

- Tree with Age attribute

Logistic Regression

Intro to Logistic Regression

- Logistic regression applications
  - Predicting the probability of a person having a heart attack
  - Predicting the mortality in injured patients
  - Predicting whether customers will purchase a product or halt a subscription
  - Predicting the probability of failure of a given process or product
  - Predicting the chance that a person defaults on a mortgage
- Which is logistic regression suitable?
  - If your data is linear.
  - LR, SVMs, Naïve Bayes
  - If you want probabilistic results
  - Naïve Bayes needs a linear decision boundary.
  - If you want to prevent the impact of features.

Logistic Regression vs Linear Regression

- Adjusted function in logistic regression
  - Logistic Function

- Classification is the extreme known model

What is the output of our model?

- P(Y=1|X)
- P(y=0|X) = 1 - P(y=1|x)

- P(Churn=1|Income,age)=0.8
- P(Churn=0|Income,age) = 1-0.8=0.2

$$e^{(\beta^T X)} \quad \rightarrow P(y=1|x)$$

$$1 - e^{(\beta^T X)} \quad \rightarrow P(y=0|x)$$

- The working process

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} \quad \rightarrow P(y=1|x)$$

1. Initialize P
2. Calculate $P \times (1-P) \times x$ for a customer
3. Calculate the output of P with actual output of customer and predicted output of customer
4. Calculate the error for all customers
5. Change that P to reduce the cost
6. Go back to step 3

$$\log(\frac{P}{1-P}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

$$\log(\frac{P}{1-P}) = \beta^T X$$

$$P = \frac{e^{\beta^T X}}{1 + e^{\beta^T X}}$$

Logistic Regression Training

- General cost function

$$cost(P, y) = \begin{cases} -\log(P) & \text{if } y = 1 \\ -\log(1-P) & \text{if } y = 0 \end{cases}$$

- Change the weight → Reduce the cost
- Cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost(P, y)$$

- Finding the cost function of the model
  - Model P
  - Actual label y(i) = 0
  - If P(i)_hat y(i) = 0 cost = 0
  - If P(i)_hat y(i) = 0 cost = large

- Logistic regression cost function
  - So, we will replace cost function with:

$$cost(P, y) = \begin{cases} -\log(P) & if\ y = 1 \\ -\log(1-P) & if\ y = 0 \end{cases}$$

$$cost(P, y) = -y\log(P) - (1-y)\log(1-P)$$

- Minimizing the cost function on the model
  - How to find the best parameters for our model?
  - Minimize the cost function
  - How to minimize the cost function?
    - Using Gradient Descent
  - What is gradient descent?
    - It is a method to calculate derivative of cost function to change the parameter values, in order to minimize the cost

- Using gradient descent to minimize the cost

Numerical Gradient Descent

- Training logistic regression
  1. Initialize the parameters randomly
  2. Feed the cost function with training data and calculate the cost
  3. Calculate the gradient of cost
  4. Update weights
  5. Update weights until new values give a acceptable cost
  6. Go to step 2
  7. Predict the new customer's

$$\theta_1 = \theta_1 - \alpha\frac{\partial J(\theta)}{\partial \theta_1}$$

$$\theta_2 = \theta_2 - \alpha\frac{\partial J(\theta)}{\partial \theta_2}$$

$$\theta_n = \theta_n - \alpha\frac{\partial J(\theta)}{\partial \theta_n}$$

$$\theta := \theta - \alpha\frac{\partial J(\theta)}{\partial \theta}$$

Support Vector Machines

Support Vector Machine

- What is SVM?
  - SVM is a supervised algorithm that classifies cases by finding a separator.
  - Mapping data to a high-dimensional feature space
  - Finding a separator.

- Data transformation

$$\phi(x) = (x^2, x^3)$$

Question

- Which of the following is/are true?
  1. Linear kernel is the most frequently used kernel.
  2. Kernel function is a similarity function.
  3. Kernel function is a dot product of two vectors.
  4. Kernel function is used to map the data into a higher dimensional space.

- Using SVM to pick the hyperplane

- Pros and cons of SVM
  - Advantages:
    - Accurate in high-dimensional spaces
    - Memory efficient
  - Disadvantages:
    - Prone to over-fitting
    - No probability estimation
    - Small datasets
- SVM applications:
  - Image recognition
  - Text category assignment
  - Detecting spam
  - Sentiment analysis
  - Gene expression classification
  - Regression, outlier detection and clustering