Machine Learning

Group 16 - Assigment 1

Luca Pfrang	5332329
Tran Luong Bang	5365615
Louis Ngo	4506205
Vu Thi Hoang Anh	5367008
Tidiane Ndir	5364532

Freiburg, October 23 2021

1. General Questions

1.1. What is main goal of machine learning

The main goal of ML is to learn to approximately solve a task using past occurrences of task instances and given solutions by maximizing the quality of the approximated solutions. ML tasks could be speech recognition, language natural processing, computer vision… and thousands of other applications.

1.2. What are different types of learning

There are 3 main methods are used today

- *Supervised Learning*
    - Training data with given label
    - Typical tasks: Classification, Regression, Raking
    - Some of most important supervised learning algorithms:
        - K- Nearest Neighbors
        - Linear Regression
        - Logistic Regression
        - Support Vector Machines
        - Decision Trees and Random Forests
        - Neural networks
- *Unsupervised Learning*
    - The training data is unlabeled
    - Some of most important unsupervised learning algorithms:
        - Clustering: K-Means, DBSCAN, Hierarchical Cluster Analysis
        - Anomaly detection and novelty detection: PCA, Kernel PCA
        - Visualization and dimensionality reduction
- *Reinforcement Learning* is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward.

1.3. How would you describe the overfitting and underfitting phenomenon.

- *Overfittin*g: Overfitting means that model performs well on the training data but it does not generalize well on unseen data.
    - There are some techniques to prevent overfitting:
        - Gather more training data
        - Apply regularizations
        - Using simple models
        - Cross Validation

- *Underfitting:* Underfitting is opposite of overfitting: it occurs when model to too simple to learn the underlying structure of the data.

- There are some techniques to reduce underfitting:
  - Increase model complexity
  - Increase the number of features
  - Remove noise from data

2. K-nearest Neighbors

2.1. Solution attached in **ex_sheet_1.py**

2.2. What is the main advantages and disadvantages of K-nearest neighbor algorithm.

- Advantages of KNN:
  - *No training Period*: KNN is called Lazy Learner. It stores training dataset and learns from it only at the time of making real time prediction. That make KNN algorithm much faster than other algorithms that require training e.g. SVM etc.
  - *Easy to implement*: There are only two parameters required to implement KNN i.e. The value of K and the distance function (Euclidean or Manhattan Distance)
- Disadvantages of KNN:
  - Computational complexity: expensive cost to calculating the distance in the large dataset or dataset with high dimensions.
  - Poor performance on imbalanced data
  - Sensitive to noisy data and missing values and outliers

2.3. Do you think the value of K affects the results? Provide a few scenarios behind your reasoning.

- Choosing K significantly affects to the performance of the model, if chosen incorrectly it can cause the model to be over or under fit. If value of K is too small, it would cause noise in data have a high influence on the prediction, however if value of K is too large, it makes computation expensive.
- There are some methods to choose the right K:
  - By taking the square root of N where N is total number of samples
  - Experiment with various values of K and their associated accuracies.

3. Naïve Bayes

Car	Color	Type	Origin	Stolen
1	red	sports	domestic	yes
2	red	sports	domestic	no
3	blue	sports	domestic	yes
4	blue	sports	domestic	no
5	blue	sports	imported	yes
6	blue	grand tourer	imported	no
7	blue	grand tourer	imported	yes

8	blue	grand tourer	domestic	no
9	red	grand tourer	imported	yes
10	red	sports	imported	yes

3.1.Estimate the probability

- P(yes) = 6/10 = 0.6
- P(red | yes) = 3/6 = 0.5
- P(grand tourer | yes) = 2/6 = 0.33
- P(domestic | yes) = 2/6 = 0.33
- P(no) = 4/10 = 0.4
- P(red | no) = 1/4 = 0.25
- P(grand tourer | no) = 2/4 = 0.5
- P(domestic | no) = 3/4 = 0.75

3.2. Predict the probability P(yes | red, grand tourer, domestic)
Assume that Color, Type and Origin are all independent given y:

P(yes | red, grand tourer, domestic) $\propto$ P(yes)*P(red | yes) *P(grand tourer | yes) *P(domestic | yes) = 0.6*0.5*0.33*0.33 = 0.033

P(no | red, grand tourer, domestic) $\propto$ P(no)*P(red | no) *P(grand tourer | no) *P(domestic | no) = 0.4*0.25*0.5*0.75 = 0.0375

So the car is less likely to be stolen.

3.3. In general: What are the benefits? What are the downsides of using Naïve Bayes?

- Benefits:
  - Very fast to train and to predict
  - Models works well with high-dimension sparse data and are relatively robust to the parameters.
  - Naïve Bayes models are great baseline models and are often used on very large datasets, where training even a linear model might take too long.
- Downsides:
  - Naïve Bayes assume that all features are independent, rarely happening in real life.
  - If categorical variable has a category in test dataset which is not observed in training dataset, then model will assign a 0 probability and unable to make a prediction. It is well-known as Zero Frequency.

3.4. The extra mile: Derive Equation 1 using Bayes' theorem, the chain rule of probabilities and the conditional independence assumption stated above.

According to Bayes' Theorem:

$$P(y = k \mid x_1, x_2, \dots, x_n) = P(y = k) \frac{P(x_1, x_2, \dots, x_n \mid y = k)}{P(x_1, x_2, \dots, x_n)} \quad (1)$$

Applying the chain rule to the numerator:

$P(x_1, x_2, \dots, x_n \mid y = k) = P(x_1 \mid y = k).P(x_2 \mid x_1, y = k)..P(x_n \mid x_1, \dots, x_{n-1}, y = k)$ (2)

Using the assumption that $x_1, x_2, \dots, x_n$ are independent (3) and $P(x_i \mid x_{i+1}, \dots, x_n, y) = P(x_i \mid y)$

RHS of equation (2) become $P(x_1, x_2, \dots, x_n \mid y = k) = \prod_{i=1}^{N} P(x_i \mid y = k)$ (4)

We have

$$Z = P(x_1, x_2, \dots, x_n) = \sum_{k=1}^{K} P(y = k)\, P(\,(x_1, x_2, \dots, x_n \mid y = k)$$
$$= \sum_{k=1}^{K} P(y = k) \prod_{i=1}^{N} P(x_i \mid y = k) \quad (5)$$

From (1) (3) (4) (5) we can write Naïve Bayes as:

$$P(y = k \mid x_1, x_2, \dots, x_n) = \frac{1}{Z} P(y = k) \prod_{i=1}^{N} P(x_i \mid y = k)$$

4. Ranking Losses

4.1. Formulate mathematically a loss function that evaluate how well some generic prediction matches $\hat{y}$ the target ranking y.

Mean Square Error is the most commonly used loss function for regression. The loss is the mean overseen data of the squared difference between true and predicted values, writing as a formula.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Where *y* is ground truth and $\hat{y}$ is predicted value

4.2. According to this mathematical formulation, which model is better at ranking?

In this case we have:

Loss of model 1:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{3}\left((1-1)^2 + (2-3)^2 + (3-2)^2\right) = 0.67$$

Loss of model 2:

$$L_2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{3}\left((1-2)^2 + (2-3)^2 + (3-7)^2\right) = 6$$

According to these results, we can say that model 1 has the better performance then model 2.

Why is the squared error problematic in this case?
First, least squared error is problematic because it penalizes magnitude change heavily.
Second, if we use squared error as our loss function, the more number sample of data, the more squared error increase. Fair enough? Absolutely not! When we increase our sample data, the error should decrease. That's a reason we used Mean Squared Error as the loss function in this case.