Applied Data Science

Xinyu Zhong
Queens' College

November 27, 2023

Contents

Abstract

Abstract of this course

1 Pre-processing Data

2 Understanding the structure of Data

- type of features: continuous, categorical
- ranges of features: [min, max], number of categories
- missing information [la/bels, features]
- discriminative power of features (redundancy)

Simpson's paradox:

Simpson's paradox is observed in probability and statistics; a trend appears in several groups of data but disappears or reverses when the groups are combined.

2.1 Adjusting data without tampering with signal

2.1.1 Expression ranges and One-hot encoding

For categorical data, we assign value by using one-hot encode which are the n unit vectors, which are equal distance from each other, from n-dimensional space to represent n categories

2.1.2 Standardisation vs scaling

There are different ways of scaling, the most straightforward one is linear scaling, others include log scaling. Note that outliers will seriously affect scaling, it can be done by cap our value, however, it also means that we throw away data points Note that we lose certain information when standardise, i.e. for Z

$$Z = \frac{x_i - \mu}{\sigma}$$

you will lose the standard deviation. Also, μ and σ can be influenced by outliers. Median and MAD(median absolute deviation)are not affected by outliers

$$Z_{\text{med}} = \frac{x_i - \text{ median}}{MAD} \tag{1}$$

$$MAD = \text{ median } (| \, x_i - \text{ median } |) \tag{2}$$

2.1.3 Near zero variance

Near zero variance means that the variable has little variance i.e. almost constant

2.1.4 Multi-collinearity

Multi-collinearity is a concept where independent variables are highly correlated. i.e. correlation coefficient = 1 We use PCA analysis to reduce the dimension of the highly correlated variable space

2.1.5 Dimensionality reduction

2.2 Engineering Model Robustness

- we draw the samples independently and identically (iid) at random from the distribution (there is no underlying structure that is present in the data)

- the sets are disjunct partitions of the original distribution (no intersection between training set and test set) the size of the validation and test sets should be comparable (if not identical)
- The validation set should be large enough to detect differences between models
- accuracy is not the only metric to measure the performance of the model
- on the test set, the error between the prediction and the actual label is the test error
- the objective function of the algorithm minimizes the test errors by parameter tuning
- Models are further evaluated for Bias and Variance (assessment of overfitting/ underfitting)

2.2.1 Validation set

k-fold validation is essentially k models ALSO use multiple k, iterations Usually

2.2.2 Confusion matrix

Definition of confusion matrix A stacked approach to deal with data with underlying substructure. choose a representation? PCA o weighted summary

2.2.3 Unbalanced data

fixed by upweighting or down sampling ***Nyquist–Shannon sampling theorem:***

Kullback-Leiber divergence (per classes or using a binning approach for continuous data) :

3 Supervised Learning: Regression

We assume the model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 and β_1 are two unknown constants that represent the intercept and the slope; β_0 and β_1 are also known as coefficients or parameters and ϵ is the error term.

Given the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict the output, $\hat{y}$ using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{y}$ indicates the prediction of Y on the basis of $X = x$. The hat symbol denotes an estimated value.

3.1 Standard errors

3.2 Hypothesis Testing

Standard errors can be used to perform hypothesis testing on the coefficients. The most common hypothesis test involves testing the null hypothesis of: H_0 : There is no relationship between X and Y H_1 : There is some relationship between X and Y Or, more formally:

$$\begin{aligned} H_0 : &\quad \beta_1 = 0 \\ H_1 : &\quad \beta_1 \neq 0 \end{aligned}$$

To test the H_0 we compute a t-statistic:

$$t = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)}$$

This will be a t-distribution with $n-2$ degrees of freedom. Using R, we can compute the probability of observing any value $\geq |t|$. We call this probability **p**-value.

3.3 Type of Loss Functions

- Mean Absolute Error (MAE) (L1 Loss)
- Mean Squared Error (MSE) (L2 Loss)
- Mean Biased Error (MBE)
- Hubber Loss (L1-L2 Loss)

3.4 Bias and Variance Trade-off

Bias: the model's error rate on the training set (rephrased, the difference between the average prediction and the correct value we are predicting).

A model with high bias is oversimplified (insufficient information acquired from the training data).

Variance: the model's error rate on the validation (or test) set, in addition to the bias

A model with high variance captures the signal and the noise in the training data and fails to generalise well on (unseen) test data.

3.5 Multiple (linear) regression. Model selection

How to choose which subsets to use, for n features, there are 2^n subsets, which is not feasible to try all of them.

3.6 Model selection.

Forward Selection Begin with the null model, a model that contains an intercept but no predictors
-Fit p simple linear regressions, each with only one feature and add to the null model the variable that results in the lowest RSS.
-Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold

backward selection Start with all variables in the model; fit the model

Remove the variable with the largest p-value i.e. the variable that is the least statistically significant

The new $p-1$-variable model is fit, and the variable with the largest p-value is removed.

Continue until a stopping rule is reached e.g. when all remaining variables have a significant p-value above some threshold

3.7 Parametric logistic regression

3.8 Non-parametric regression

Decision tree, where the data space is partitioned into regions

The complexity parameter prevents the tree from having too many branches, which is prone to overfitting

3.9 Regularisation

Regularisation is the process of adjusting an algorithm to prefer a smaller model, to avoid overfitting. This is done by modifiying the loss function to include a penalty for large weights.

4 Machine Learning Model - KNN

K nearest neighbour

- KNN is a non-parametric method
- KNN distance can be defined with Eucleadian distance, Manhattan distance, etc.s
- Lower K means a more complex model (Prone to overfitting)
- Higher K means a less complex model (Prone to underfitting). In the most extreme case, K = N, then your model would predict the same outcome for all data points.

4.1 Standardisation

Standardisation is necessary as we need to make sure distances are equally weighted. Another method is to use correlation-based distance. Check the skewness before Standardisation

5 Machine Learning Model Support Vector Machine

There is always a misclassification Transform points to higher dimensions to allow linear separation

6 Machine Learning Model - Decision Tree

A decision tree is a non-parametric method, therefore it does not require data pre-processing.

6.1 Recursive binary partitions

Resoureces: An Introduction to Statistical learning with Applications in R/Python Segment feature space into simple regions (high demension rectangles) Predict usinging the average, mode, classify using the most common class

6.2 Characteristic of a decision tree

- branches: represent the decision rules
- interial nodes:
- leaf nodes (terminal nodes): represent the outcome,

6.3 How to grow a tree

How to grow the tree? - General process: 1. Partition the predictor space (i.e. all possible values of $X_1, X_2, \ldots, X_p$) into J distinct, non-overlapping regions, labelled $R_1, \ldots, R_J$ 2. Each observation in a given region R_f is given the same predicted value (or class), which is the mean (or mode) of all response variables in that region - How do we choose the partitions? e.g. high-dimensional rectangles

Greedy approach - regression regression trees - For one of the predictors X_j with a cutoff threshold c, we define two regions:

$$R_1(j, c) = \{X \mid X_j < c\} \text{ and } R_2(j, c) = \{X \mid X_j \geq c\}$$

- An observation x_i is in region $R_1(j, c)$ if $x_i < c$ and in region $R_2(j, c)$ if $x_{ij} \geq c$ - For the first iteration, $R_1(j, c)$ and $R_2(j, c)$ partitions the entire predictor space - For tater tterations, $R_1(j, c)$ and $R_2(j, c)$ partitions a previous parent' region (so inherits any previous conditions of that 'parent' region)

Classification The function to minimise is to reduce the Gini-index,

6.4 Bagging

Bagging is the process of creating multiple models from different subsets of the training data, of the same size, by bootstraping. The final prediction is averaged across (majority vote or average across) the predictions of all of the sub-models.

6.5 Random Forest

Random forest is an ensemble method that combines multiple decision trees. It is a bagging method with a random selection of features. The final prediction is averaged across the predictions of all of the sub-models.

6.6 Boosting

Boosting sequentially builds a model by training a decision tree on the residuals of the previous model. The final prediction is the sum of the predictions of all of the sub-models. The parameters include the learning rate, which is the shrinkage factor of the model for the residuals.

7 Machine Learning Model - K-MEAN

K-mean is a clustering method, which is an unsupervised learning method. It is used to group data points into clusters, where the data points in the same cluster are similar to each other, and the data points in different clusters are dissimilar to each other.

K-mean clustering are widely used in pattern recognition, computer vision, etc.

K-mean is similar to k-nearest neighbour.

K-mean is a NP-hard problem, which means that it is computationally expensive to find the optimal solution. Therefore, we use a heuristic approach to find a sub-optimal solution.

7.1 Algorithm

7.1.1 Lloyd's algorithm (naive k-means)

Steps of Lloyd's algorithm:

1. Initialisation: Randomly assign each data point to one of the k clusters

2. Assign each data point to the closest centroid
3. Compute the centroid of each cluster
4. Repeat steps 2 and 3 until the centroids do not change

This algorithm should reduce the within-cluster sum of squares (WCSS) at each iteration.

7.1.2 K-means alternatives

- K-median: use the median instead of the mean to compute the centroid
- K-medoids: use the most central point in the cluster as the centroid
- Manhattan distance: use Manhattan distance instead of Euclidean distance to compute the distance between a data point and a centroid (i.e. $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$)

7.1.3 Initialisation

There are serveral ways to initialise the centroids:

- Random partition: Randomly assign each data point to one of the k clusters
- Forgy: Randomly select k data points as the initial centroids
- K-means++: select the first centroid randomly, then select the next centroid from the remaining data points with a probability proportional to the distance from the previous centroid

7.2 Fuzzy k-means

1. Multiple cluster assignment: x_i has cluster assignment w_{ik}, with $w_{ik}^{-1} = \sum_{j=1}^{K} \left(\frac{\|x_i - c_k\|}{\|x_i - c_j\|} \right)^{\frac{2}{m-1}}$
2. Centroid update - New cluster centroids are $c_k = \frac{\sum_{i=1}^{n} w_{ik}^{m} x_i}{\sum_{i=1}^{n} w_{ik}^{m}}$ - This minimises the weighted mean squared error $E = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{m} \|x_i - c_k\|^2$ - m is a fuzziness hyper parameter

7.3 Types of clustering

- Hard vs soft clustering
- Hierarchical vs non-Hierarchical clustering
- AgglomeraCve vs partitioning/divisive
- Centroid vs distribuCon-based vs density vs graph-based vs spectral

8 Clustering

8.1 GMM

8.2 Density

8.3 Evaluation

The challenge in evaluating clustering is that quantitative measures are not available. There is internal evaluation and external evaluation.

9 Python Library

9.1 statsmodels

statsmodels.api https://www.statsmodels.org/stable/examples/index.html

9.2 sklearn

neighbours.KNeighborsClassifier Parameters:

- n_neighbors: number of neighbours
- weights: uniform, distance
- algorithm: auto, ball_tree, kd_tree, brute. Trees reduces the complexity from $O(N)$ to $O(logN)$
- leaf_size: leaf size for ball_tree or kd_tree
- p: power in minkowski distance

output conf_matrix = confusion_matrix(y_test, y_pred) Cohen's Kappa: $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o is the accuracy and p_e is the expected accuracy, we want to maximum Kappa. The definition of p_e is:

$$P_e = \frac{\sum \left(\frac{\text{Row sum for category } i \times \text{ Column sum for category } i}{\text{Total number of observations}} \right)}{\text{Total number of observations}}$$

To simplify P_e, we can rearrange the terms:

$$P_e = \frac{\sum(\text{ Row sum for category } i \times \text{ Column sum for category } i)}{(\text{ Total number of observations })^2}$$

neighbours.KNeighborsRegressor