Applied Data Science

Xinyu Zhong
Queens' College

October 24, 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 straight forward 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 certian information when standardise, i.e. for Z

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

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

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

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

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 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 partioned into regions and the patrition

Complexity parameter prevents the tree to have too many branches, which is prone to overfitting

3.9 Regularisation

Regularisaition is th 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 Python Library

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