Model Evaluation and Validation

1 Introduction

The ultimate goal of machine learning is to have data models that can learn and improve over time. Essentially making inferences based on previous examples.

2 Measures of Central Tendency

There are three measures of central tendency:

- **Mean**: The average of the distribution
- **Median**: The value at the middle of the distribution
- **Mode**: The value at which the frequency is highest of the distribution

The mode's primary properties are that:

- The mode can be used to describe quantative and qualitative data both
- All scores in the database do not affect mode
- Given a large number of samples from the population, the mode will be not the same in each
- There is no equation for the mode

The mean's primary properties are that:

- All scores in the distribution affect the mean
- The mean can be described with a formula
- Many samples from the same population will have similar means

- The mean of a sample can be used to make inferences about the population it came from
- The mean will change if an extreme value was added

Data needs to be sorted for the median to be useful. A normal distribution is one where the median is equal to the mean and mode

3 Variability of Data

Range is a measure of the spread of the data and represents the difference between the lowest and highest values. Drastically affected by outliers.

Quartiles are used to chop remove the 1st and 4th quartiles (top 25% and bottom %) to help normalize the data.

The Inner Quartile Range (IQR) and is defined as $IQR = Q_1 - Q_2$

Outliers are extreme data values such that an $O < Q_1 - 1.5(IQR)$ or $O > Q_3 + 1.5(IQR)$

Box plots can represent the data with quartile ranges, min/max and outliers represented.

Deviation is the $\frac{\sum_i^n (\bar{x} - x_i)}{n}$ and doesn't work in measuring variability since negative values can cancel out the positive results.

Squaring or absolute value can be used to curtail that issue. Using absolute value we get $\frac{\sum_i^n (|\bar{x} - x_i|)}{n}$ or using squaring we have $\frac{\sum_i^n (\bar{x} - x_i)^2}{n}$. The latter formula is known as the variance.

The standard deviation would be defined as $\sqrt(\frac{\sum_i^n (\bar{x} - x_i)}{n})$

The normal distribution's structure depends on the standard deviation

Due to issues with population size vs sample size. If we take samples smaller than the entire population it is entirely possible to obtain situations where the sample size standard deviation does not represent the whole population standard deviation.

To correct for this we use Bessel's Correction whereby the standard deviation is $\sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}}$ and the variance is $\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$ for the sample standard deviation and the sample variance.

The sample standard deviation is denoted s whereby the population standard deviation is σ

4 Numpy and Pandas

Numpy deals with linear algebra while Pandas allows us to deal with data frames

5 Introduction to SciKit-Learn

General process is:

1. Import
2. Create
3. Fit
4. Predict

6 Nature of Data

There are several types of data

- Numeric data (quantitative data) of consists of two types: discrete and continuous where discrete data can only have one of several finite choices whereby continuous data can be chosen from any point within the a given bound
- Categorical data represents characteristics and can take on numerical values and the data doesn't have a mathematical data
- Ordinal data represents data that is not necessarily heavily mathematically describable but is order-able
- Time series data is data that is collected over time, whereby time implies the ordering methodology
- Text words, which has an incredibly complex relationship with one another, usuaully not mathematically effible

Sk-learn is capable of encoding data to prevent issues w/ sk-learn surrounding categorical or ordinal data.

7 Datasets and Questions

There is a strong correlation between accuracy and training set size suxh that the rule of diminishing returns still applies.

8 Training and Testing

Separating traininng and testing data is important as it can prevent overfitting and gives us data to test our models on.

Features are measurable properties that can be used to make predictions about labels. Features are the category and the label is the specific.

Cross validation splits a dataset up into several small subgroups and can use and recombine those groups to find the optimal training/test set ratio and combination. It runs K different experiments and then trains and tests, averaging the results from k experiments.

The kfold algorithm in Sklearn does not split data randomly an simply partitions it.

The grid search algorithm can cross validate a number of different parameter tunes to figure out which split gives the best performance for the algorithm

9 Evaluation Metrics

Cllassification is about deciding what categories things belong to. Regression is about making predictions on continuous data. That is, classification is discrete while regression is continous.

Different metrics can be used to tune algorithm performance.

There are several classification metrics that can be used:

- Accuracy: The proportion of items classfied/labeled correctly. Uses the score metric. Accuracy can't deal well with skewed classes and is quite neutral in its evaluation.
- Precision: Is the proportion of true positives to the sum of true positive and false negative.
- Recall: Is the proportion of true positive to the sum of true positive and false positive.
- Confusion Matrix is a matrix storing the main components false negatives and false positives that are used in the calculation of prediciton and recall.
- F1 Score is a combined metric of precision and recall calulcated by $\frac{precision \times recall}{precision + recall}$

A confusion matrix takes the labels and puts them in a n^2-degree matrix where n is the number of available labels for the test feature. In the confusion matrix elements are placed where they actually belong vs where they were classified.

Differing metrics are used to handle differing wieghting schemes for given features

There are several core accuracy metrics used for regressions and they are

- Mean Absolute Error: Takes the absolute error from each example and averages them.
- Mean Squared Error Is the most common metric to use. Emphasizes larger erros over smaller errors and is differentiable allowing to calculate min/max while being more computationally efficient.
- R2 Score cwhich computes coefficients of determination of prediction for true values.

10 Causes of Error

There are two major causes of error when measuring model performance:

- Bias which is a model being unable to represent the coomplexity of the data. It underfits the data. The model is inadequate.
- Variance which is a model being overly sensitive to the data. It overfits the data. Too much variance indicates that an algorithm has not been generalized properly.

We can overcome issues with bias by using an intenionally more complex model. We can overcome issues with variance by using an intentionally less complex model or by training on more data.

The learning curve function is sklearn allows us to measure the efficiency with which the given classifier/regressor is learning from any number of datapoints.

The quest to improve the validity of a model inclues minimizing the biance and variance.

There is a formula to express the tradeoff in the variance and the bias

$$Err(x) = E[(Y - \hat{f}(x))^2] = (E[]\hat{f}(x)] - f(x))^2 - E[(\hat{f}(x) - E[\hat{f}(x)])^2] + \sigma^2$$

This expression shows that you can reduce the variance or error to a point but there is some irreducible error σ that cannot.

High bias implies high error on the training set. High variance implies much higher error on the test set than the training set.

Using few features results in a higher bias scheme. Using more features results in a high variance scheme.

We want to optimize for fewer features and a large R^2 or low SSE.

11 Curse of Dimensionality

As the number of features/dimensions grows, the amount of data required to generalize accurately grows exponentially.