Notes on Classification

Sameer Kesava PhD

Contents

Improving the Accuracy

A robust and accurate model can be achieved by tackling the **bias-variance** problem.

$$
\begin{aligned}
Bias &= <\hat{f}(X)> - f(X) \qquad (1.1)\\
Variance &= <\hat{f}(X)^2> - <\hat{f}(X)>^2 \qquad (1.2)\\
<[f(X) - \hat{f}(X)]^2> &= Bias^2 + Variance \qquad (1.3)
\end{aligned}
$$

1.1 Weights

1. Sample or Class weights can be used to improve the prediction accuracy, e.g. assigning large penalties to wrong predictions of the minority class using the 'class weight' parameter.
2. Alternatively, a user-defined function can be supplied to compute the weights.
3. Weighted micro and macro-averaging useful for scoring imbalanced datasets.

1.1.1 ML

1.1.2 DL

1. A good starting values for the initializers are Glorot-Xavier or Orthogonal for the weights.
2. For the bias, zeros suffice.

1.2 Cross Validation

- Can also use ***Bootstrap***.
- Once the best model is chosen through CV, then the model should be trained on the whole dataset for the final model.

1. Use k-fold CV. k = 10 is a good starting value.
2. Use stratified k-fold to preserve class proportions.

1.3 Regularization

1.3.1 LASSO or L1-norm

A useful feature selection technique

1.3.2 Ridge Regression or L2-norm

1.3.3 L1-L2 norm

- Implemented as ElasticNet in scikit-learn.

1.4 Brute-Force

Brute-force algorithm is effective for small datasets.
Query time grows as O[pn], where p is the number of features/predictors/dimensions and n is no. of data points.

1.5 Small Dataset

Also see Augmenting Data section 1.8.
n: number of samples
p: number of features/predictors
When p > n, then there is no unique solution - variance is ∞.

1.5.1 When p is comparable to n

1. Use shrinkage methods such as Lasso or L1 norm with k-fold CV for feature selection.

2. Check Brute-Force section 1.4.

1.5.2 $p \ll n$

1. Use k-fold CV.

1.6 Class Imbalance

Class imbalances can be handled using the following approaches.

- Upsampling.

- Downsampling.

- Generation of synthetic training samples using methods such as SMOTE.

- Data augmentation; see section 1.8.

1.7 Ensemble Learning

Random Forest, for example, is a type of ensemble learning method as different trees train on different subsets of the samples(?).

- Use Nested CV for ensemble learning.

- Boosting

- Bagging

- Stacking

- Majority/Plurality voting for multiclass setting.

- Cloning estimators (sklearn.base import clone) for fitting on different formats of training data.

1.7.1 Bagging

- Used in Random Forests, random feature subsets are selected with replacement.
- Improves accuracy of unstable models.
- Reduces overfitting by reducing variance.
- Ineffective in reducing model bias. Hence, should be used in conjunction with ensemble classifiers with low bias such as unpruned decision trees, i.e. where max_depth = 0.

1.7.2 Boosting

- Adaptive Boosting (from sklearn.ensemble import AdaBoostClassifier)
  1. Focus on samples that are hard to classify.
  2. Can lead to decrease in both bias and variance.
  3. Uses an ensemble of classifiers: the i+1th learner has it's input the output of the ith learner.

1.7.3 Stacking

- For reducing bias.
- Uses different learners and then a meta-learner in the end which takes the output of different learners to yield the response.

1.8 Augmenting Data

Accuracy can be improved by augmenting to the data and then fitting the model with both the original and augmented datasets. Some examples are

- Adding random noise to the data.
- In case of images, random rotations of the images.

1.9 Hyperparameter Tuning

- Hyperparameter optimization based on Gaussian Process Regression and Bayesian Optimization
- keras tuner in keras
- GridSearchCV or RandomSearchCV in scikit-learn
- RandomSearch can be used as the baseline against which optimization algorithms can be evaluated.

1.10 Step Functions

- Instead of a single function fitting to all the data, i.e. global fitting, step function approach allows local fitting.
- Popular in biostatistics and epidemiology.
- Wavelet or Fourier series can be used to construct basis functions in spline fitting.
- CV can be used to determine the number of knots.
- RandomForest regression is like spline fitting, i.e. it is a sum of piecewise functions (linear in this case).

ML Methods

2.1 Linear Regression

Can be used as the ***base model***.

2.2 Principal Coefficient Regression

- Works well when the variance can be explained by the first few principal components.
- Disadvantage is that PCR does not account for response variable Y.

2.3 Partial Least Squares

- In contrast to PCR, uses response variable Y.

2.4 Generative Additive Models

- For multivariate data, GAMs can be used for regression.
- For more general models, random forest or xgboost regressors should be explored.

DL Methods

All operations, e.g. loss functions, activation functions, in neural networks must be differentiable.

Loss Functions

4.1 MAE

- Can be used if there are not many outliers present.

4.2 MSE

4.3 MSLE

- Mean-Squared Log Error
- Used when Y is large.

Statistical Tests

5.1 T-statistic

When you run a hypothesis test, you use the T statistic with a p value.

$$t - statistic = \frac{\hat{\beta}}{SE(\hat{\beta})} \tag{5.1}$$

$\hat{\beta}$ is the coefficient estimate from the fitting.

5.2 Z-statistic or Z-score

statisticshowto.com: Simply put, a z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically its a measure of how many standard deviations below or above the population mean a raw score is.
A z-score can be placed on a normal distribution curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). In order to use a z-score, you need to know the mean and also the population standard deviation .
Z-scores are a way to compare results to a normal population. Results from tests or surveys have thousands of possible results and units; those results can often seem meaningless. For example, knowing that someones weight is 150 pounds might be good information, but if you want to compare it to the average persons weight, looking at a vast table of data can be overwhelming (especially if some weights are recorded in kilograms). A z-score can tell you where that persons weight is compared to the average populations mean weight.

5.3 Confidence Interval

For normally distributed error, 95% confidence interval corresponds to 2σ.