Applied Predictive Modeling Notes

Judah

04/17/24

Contents

1 Introduction **2**
1.1 Prediction Versus Interpretation 2
1.2 Ingredients of Predictive Models 2
1.3 Terminology 3
1.4 Notation 3
1.5 Other Notational Guidelines 4

I General Strategies **4**

2 A Tour of the Predictive Modeling Process **4**
2.1 The Chapter's Case Study: Predicting Fuel Economy 4
2.2 Themes 6

3 Data Pre-processing **7**
3.1 The Chapter's Case Study: Cell Segmentation in High-Content Screening 7
3.2 Data Transformations for Individual Predictors 7
3.3 Data Transformations for Multiple Predictors 9
3.4 Dealing with Missing Values 9
3.5 Removing Predictors 9
3.6 Adding Predictors 9
3.7 Binning Predictors 9
3.8 Computing 9

4 Over-Fitting and Model Tuning **9**

1 Introduction

Definition 1 (Predictive Modeling)**.** *The process of developing a mathematical tool or model that generates an accurate prediction.*

Why do predictive models fail?

1. Inadequate pre-processing of the data.
2. Inadequate model validation.
3. Unjustified extrapolation, e.g.:the application of the model -> data is in an application not seen by the model.
4. Over-fitting the model to the existing data.

1.1 Prediction Versus Interpretation

For applications that require accurate predictions:

- Historical Data — Interested in accurately projecting the chances of future events, not why they occurred.
- Pricing Homes on Zillow — Interested in accurate price estimates, not how they predicted them.
- Medical — Predict patient response to a treatment based on a significant number of factors.

Secondary considerations are given to the interpretation of the data, not primary. Higher accuracy models are more complex, and as a consequence reduce interpretability.

1.2 Ingredients of Predictive Models

- The best models are influenced and produced by modelers with expert knowledge and field context of the problem.
- These modelers can pre-filter irrelevant information and place more meaningful constraints on data sets.
- Modelers can also put personal biases on data.
- Predictive modeling is not a substitute for intuition, but a complement.
- Traditional experts make better decisions with results of statistical prediction.

1.3 Terminology

Definition 2 (Sample, Data Point, Observation, Instance)**.** *A single independent unit of data (A Customer), or a subset of data points (A Training Sample).*

Definition 3 (Training Set)**.** *Contains the data used to develop models.*

Definition 4 (Validation Set)**.** *Contains the data used to evaluate the performance of the final set of candidate models.*

Definition 5 (Predictors, independent Variables, Atrtributes, Descriptors)**.** *The data used as input for the prediction equation.*

Definition 6 (Outcome, Dependent Variables, Target, Class, Response)**.** *The outcome event, or quantity that is being predicted.*

Definition 7 (Continuous Data)**.** *Has natural, numeric scales. Ex: Blood pressure, cost, quantity.*

Definition 8 (Categorical, Nominal, Attribute, or Descrete Data)**.** *Has specific values without scale.*

Definition 9 (Model-Building, -Training, Parameter Estimation)**.** *The process of using data to determine values of model equations.*

1.4 Notation

1. n = the number of data points.
2. P = the number of predictors.
3. y_i = the ith observed value of the outcome, $i = 1 \dots n$.
4. $\hat{y}_i$ = the predicted outcome of the ith data point, $i = 1 \dots n$.
5. $\overline{y}$ = the average or sample mean of the n observed values of the outcome.
6. $\mathbf{y}$ = a vector of all n outcome values.
7. x_{ij} = the value of the jth predictor for the ith data point, $i = 1 \dots n$ and $j = 1 \dots P$.
8. $\mathbf{x}_i$ = a collection of the P predictors for the ith data point, $i = 1 \dots n$.
9. $\mathbf{X}$ = a matrix of P predictors for all data points; this matrix has n rows and P columns.
10. $\mathbf{X}'$ = the transpose of $\mathbf{X}$; this matrix has P rows and n columns.

1.5 Other Notational Guidelines

1. C = the number of classes in a categorical outcome.
2. C_l = the value of the lth class level.
3. p = the probability of an event.
4. p_l = the probability of the lth event.
5. $P_r[.]$ = the probability of event.
6. $\sum_{i=1}^{n}$ = the summation operator over the index i.
7. $\mathbf{\Sigma}$ = the theoretical covariance matrix.
8. $E[\cdot]$ = the expected value of $[\cdot]$.
9. $f(\cdot)$ = a function of $[\cdot]$; $g(\cdot)$ and $h(\cdot)$ also represent functions throughout the text.
10. β = an unknown or theoretical model coefficient.
11. b = an estimated model coefficient based on a sample of data points.

Part I
General Strategies

2 A Tour of the Predictive Modeling Process

2.1 The Chapter's Case Study: Predicting Fuel Economy

Contextual Notes From the Reading

- Website: fueleconomy.gov
- Dataset records vehicle characteristics like:
  - Engine Displacement
  - Number of cylinders
  - Laboratory Measurements of MPG

- etc ...

- This example focuses on high-level design principles:
  - Uses a single predictor (engine displacement), and a single response (unadjusted highway MPG) for 2010 to 2011 model year cars.

- Steps to Solve the Case Study:
  1. Understand the Data. This case uses graphs to visualize displacement to efficiency.
  2. Build and Evaluate a Model. The standard approach is to take random samples; but in this case we can use the 2010 data as the training set, and the 2011 data as the validation set.
  3. Measure Performance. For regression problems that predict a numeric, the residual data contains important information.
     - Residuals are computed as the observed value minus the predicted value: $y_i - \hat{y}_i$.
     - The root mean square error (RMSE) is commonly used to evaluate models. This is evaluated as how far Residuals are from zero.
  4. Define Relationships Between Predictor and Outcome.
     - The training set is used to estimate values needed by the model.
     - The test set is used once a few strong candidate models have been finalized. If used too much beforehand, it negates its utility.
  5. Resampling. Different subversions of the training data set are used to fit the model. In this model, 10-fold cross-validation was used to estimate the RMSE
  6. Trying Other Models. In this case we try a quadratic model that contains a squared term. Quadratic models can perform poorly on the extremes of a data set.
     - Another model is the multivariate adaptive regression spline (MARS).
     - MARS fits separate linear regression lines for different ranges.
     - MARS has a, 'tuning parameter' which cannot be directly estimated from the data.

- MARS allows up to five model terms.

7. The lowest RMSE value associated with four terms of the MARS model was chosen as the best predictor.

2.2 Themes

Data Splitting How we allocate data to tasks. In this example, we tested how the data extrapolates to a different population. If we wanted to predict from the same population, instead of a new population, a random sample would be more appropriate using interpolation.

Predictor Data This example used engine displacement as a predictor. We should use as many predictors as possible when building models. Look at how the case study was unable to predict fuel economy for engines with low displacement. More predictors could help us dial in these values with better accuracy.

Estimating Performance We used RMSE for quantitative assessment and resampling to help the user understand how techniques perform on the dataset. We used visualizations of a model to discover how it performed.

Evauluating Several Methods We evaluated three different models for our data. Without having substantive information about the problem, no single method could be said to perform better than another. It is always recommended to try multiple, widely variable techniques; then limit the choices based on the results.

Model Selection There are different types of model selection. In this case, we chose specific models, like linear regression, and also chose a type of a model, like the MARS parameters. In both cases, cross validation produced a quantitative assessment which allowed us to further filter out model choices.

Summary The process seems simple: Pick a model technique, plug in data, generate a prediction. Although this produces a predictive model, it most likely will not produce a reliable, and trustworthy model for new samples. We must understand both the data, and the objective in creating the model in the first place. We then must pre-process and select relevant data for our models.

3 Data Pre-processing

Different models have different sensitivities to predictor types. How the predictors are entered into the model matter. Data pre-processing is determined by the model. For instance, tree-based models are notably insensitive to the characteristics of predictor data, and opposite of linear regression models which are very restricted.

Definition 10 ((Un-)Supervised Data Processing)**.** *The outcome variable is (or isn't) considered by the pre-processing techniques.*

Definition 11 (Feature Engineering)**.** *The process of how predictors are encoded.*

Examples of Date Encoding

- The number of days from a reference date.
- Isolating the day, day of week, month, year as separate predictors.
- The numeric day of the year (ignoring the calendar year).
- Whether the date was within the school year (opposed to holidays or summer sessions).

3.1 The Chapter's Case Study: Cell Segmentation in High-Content Screening

Contextual Notes From the Reading

- Examination of a sample under a microscope to assess the desired outcome.
- Automated, high-throughput assessment of cell characteristics can produce misleading results.
- Issues related to segmentation of cells resulted in false-positive reading of cell damage.

3.2 Data Transformations for Individual Predictors

Transformations of predictor variables can be used for modeling techniques with strict requirements, or for difficult data used in specific models. The following discusses transformations used in this chapter's case study.

Centering and Scaling

Definition 12 (Centering)**.** *Centering the predictor set is the result of: taking averaging the predictor set, and subtracting this value from all predictor variables. The resulting mean of the set of predictor variables is zero. Let* $\boldsymbol{x}_i = \{x_1, x_2, \ldots, a_n\}$ *be the set of predictor variables, then the centered set of predictor variables is equal to:*

$$S = A - \frac{\sum_{i=1}^{n} \{\boldsymbol{x}_i\}}{n}$$

Definition 13 (Scaling)**.** *Scaling the predictor set is a result of dividing each variable by the standard deviation. This forces all variables to have a common scale, and improve numerical stability of some calculations. The only downside could be a loss of interpretability since the data is no longer in it's original unit.*

Transformations to Resolve Skewness

Definition 14 (Skewness)**.** *The* ***skew*** *of a distribution tells the bias-location of the dataset. For example, right skewed is biased towards larger values, and un-skewed is roughly symmetric. Let* $\boldsymbol{x}_i = \{x_1, x_2, \ldots, x_n\}$*, then the* ***skewness*** *of the set is:*

$$skewness = \frac{\sum (\boldsymbol{x}_i - \bar{x})^3}{(n-1)v^{3/2}},$$

where

$$v = \frac{\sum (\boldsymbol{x}_i - \bar{x})^2}{(n-1)}.$$

You can adjust the skew of a data set by applying a transformation with an equation; ex: log, ln, sqrt, inverse, etc ... You could also have a system of transformations on different ranges of data. Box and Cox Transformation Family

$$x^* = \begin{cases} \frac{x^\lambda}{\lambda} & \text{if } \lambda \neq 0 \\ \log x \& \text{ if } \lambda = 0 \\ \frac{x^\lambda}{\lambda} \end{cases}$$

3.3 Data Transformations for Multiple Predictors

3.4 Dealing with Missing Values

3.5 Removing Predictors

3.6 Adding Predictors

3.7 Binning Predictors

3.8 Computing

4 Over-Fitting and Model Tuning

We can make up some more text that is different.

text