Statistics

Cross Validation

1. Wikipedia: Cross-validation, sometimes called rotation estimation,[1][2][3] is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.
2. When the same data are used both to estimate parameters and to assess fit, there is a strong tendency towards overfitting. Overfitting means selecting an unnecessarily complex model to fit the noise. The obvious remedy is to test model fit using dat that are independent of the data used for parameter estimation.
3. A common criterion for judging fit is the deviance, which is -2 times the log-likelihood:

$$-2 \log f(Y^{val}|\hat{\theta}^{train})$$

, where $\hat{\theta}^{train}$ is the MLE of the training data and Y^{val}.

4. When the sample size is small, splitting the data once into training and calidation sata is wasteful. A better technique is cross-validation. K-fold cross-validation divides the data set into K subsets of roughly, equal size. Validation is done K times. IN the kth validation, $k = 1, ..., K$ the kth subset is the validaiton data and the other $K - 1$ subsets are combined to form the traingin data. The K estimates of goodness-of-fit are combined, for example by averaging them. A common choice is n-fold cross-validation, also called leave-one-out cross-validation. With leave-one-out cross-validation each observation takes a turn at being the calidation data set, with the other n-1 observations as the training data.
5. Akaike Information Criterion (AIC): An alternative to actually using validation data is to calculate what would happen if new data could be obtained and used for validation. AIC is an approximation to the expected deviance of a pypothetical new sample that is independent of the actual data:

$$AIC = -2 \log f\{Y^{obs}|(\hat{Y^{obs}})\} + 2p$$

6. Back-testing. Traders usually developo trading strategies using a set of historial data and then test the strategies on new data.

Different types of seasonality in time series.

Logistic Regression

1. Describe a recent use of logistic regression. # Ridge regression

maximum Likelihood

1. Bayesian inference vs MLE
2. $\hat{\sigma}^2_{MLE} = n^{-1} \sum_{i=1}^{n} (Y_i - x_i^T \hat{\beta})^2$ - is biased downwards, but the bias can be eleminated if $n^{-1}$is replaced by $\{n - (p+1)\}^{-1}$ # Box Cox transformation

- $$Y^{(\alpha)} = \beta_0 + X_{i,1}\beta_1 + ... + X_{i,p}\beta_p + \epsilon_i$$

  where $\epsilon_1, ..., \epsilon_n$ are i.i.d. $N(0, \sigma_\epsilon^2)$ for some σ_ϵ In contrast to the TBS (transform both sides) model only the response is transformed.

- objectives:
  - a simple model: $Y_i^{(\alpha)}$ is linear in predictors $X_{i,1}, ..., X_{i,p}$ and in the parameters $\beta_1, ...\beta_p$
  - constant residual variance
  - Gaussian noise

- Box and Cox suggest estimation of α by Maximum likelihood

K-means

1. How to define the number or clusters (Elbow curve)

Machine Learning Algos

1. Decision Trees
2. Neural Networks (back propagation)
3. Probabilistic networks
4. Nearest Neighbor
5. Support vector machines

Machine Learnign techniques

1. Supervised Learning
2. Unsupervised Learning
3. Semi-superivsed Learning
4. Reinforcement Learning
5. Transduction
6. Learnign to learn

Regularization

Randomization in experimental design

Multicollinearity

1. THe variance inflation factor(VIF) of a variable tells us how much the SSE is increased by having the other predictor variables in the model. For example, if a variable has a VIF of 4, then the varianceo of its $\hat{\beta}$ is four times larger than it would be if the other predictors were either deleted or were not correlated with it. The SE is increased by a factor of 2.

- Suppose we have predictor variables $X_1, ..., X_p$. Then VIF of X_j is found by regressing X_j on the $p-1$ other predictors. Let R_j^2 be the R^2 - value ot this regression, so that R_j^2 measures how well X_j can be predicted from the other Xs. Then VIF of X_j is

$$VIF_j = \frac{1}{1 - r_j^2}$$

- When interpreting VIFs, it is important to keep in mind that VIF_j tells us nothing about the relationship between the response and jth predictor. Rather, it tells us only how correlated the jth predictor is with the other predictors. In fact, the VIFs can be computerd wothout knowing the calues of the response variable.
- The usual remedy to collinearity is to reduce the number of predictor variables by using one of the model selection criteria (AIC, BIC…).
- VIF values give information about liner relationships between the predictor variables, but not about their relationships with the response.

Regression: Troubleshooting

1. Leverage:

   - a high leverage point is not necessarily a problem, only a potential problem.
   - the high-leverage point has a high influence on the estimated slope.
   - $H_{ii} - measures how much influence Y_i has on its own fitted value $\hat{Y}_i$.

   $$\hat{Y}_i = \sum_{j=1}^{n} H_{ij} Y_j$$

   . H_{ii} is the weight of Y_i in the determination of $\hat{Y}_i$. It is a potential problem if H_{ii} is large since then $\hat{Y}_i$ is determined too much by Y_i itself and not enough by the other data.
   - The leverage value H_{ii} is large when the predictor variables for the ith case atypical of those values in the data, for example, because one of the predictor variables for that case is extremely outlying.

2. Cook's Distance (Cook's D)

   - measures how much the fitted values change if the ith observation is deleted.
   - Let $Y_j(\hat{-i})$ be the jth fitted value using estimates of the $\hat{\beta}$ obtained with the ith observation deleted. Then Cook's D for the ith observation is

   $$\frac{\sum_{j=1}^{n} \{\hat{Y}_j - \hat{Y}_j(-i)\}^2}{(p+1)s^2}$$

   - One way to use Cook's D is to plot the values of Cook's D against case number and look for unusually large values.

Course of dimensionality

Naive Bayes

1. conjugate-prior

Selectin bias

R^2

Eucledian Distance of a dataset

A/B test

finate precision in machine learning

Bias variance trade-off

Data cleaning techniques

model lift

Cointegration

1. Two time series, $Y_{1,t}$ and $Y_{2,t}$ are cointegrated if each is $I(1)$ but if there esists λ s.t. $Y_{1,t} - \lambda Y_{2,t}$ is stationary.
2. The Phillips - Ouliaris conitegration test regresses one integrated series on others applies the phillips-Perron unit root test to the residuals. The null hypothesis is that the residuals are unit root nonstationary, whihc implies that the series are not conitegrated. Therefore, a small p-value implies that the series are cointegrated and therefore suitable for regression analysis.

Resources:

1. Wikipedia
2. Ruppert. Statistics for Data Analysis for Financial Engineering