DSCI 552 Cheat Sheet: Katie Foss

Lesson 1: Introduction

Unsupervised Learning: Cluster Analysis, Dimensionality Reduction (K-means, PCS, ICS)
No Free Lunch Theorem: All algorithms perform equally when averaged over all possible problems
Bias-Variance Trade-Off: As flexibility of $\hat{f}$ $\uparrow$, then variance $\uparrow$ and bias $\downarrow$. Choose flexibility based on average test error. Bias is connected to the training error.
K-Nearest Neighbors Classifier: Find K neighbors closest (from training) to new observation, majority vote. Distance metric can be euclidean or other metric. Computationally expensive (compute distance to all known samples).
- KNN performs poorly with large p because curse of dimensionality. C(x) may still work, but probabilities are off.
- Classes do not have to be linearly separable.
- Sensative to imbalanced datasets and irrelevant inputs.

$\downarrow k \rightarrow \uparrow variance\ and\ \downarrow bias$
$\uparrow k \rightarrow \downarrow variance\ and\ \uparrow bias$
KNN: $\hat{p_k}(x) = \frac{\#\ class\ k\ samples \in N(x)}{\#\ samples \in N(x)}$
$Bias[\hat{f}(x_0)] = E[\hat{f}(x_0)] - f(x_0)$
MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$
Normalization: $x' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$
Standardize: $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x_j}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}}$
Normal PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
Sample Variance: $S^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$

Lesson 2: Linear Regression

Linear Regression: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
If $n > 30$ and $\epsilon \sim N(0, \sigma^2)$, then $\hat{\beta}_0$,$\hat{\beta}_1 \sim N(\beta_1, \sigma^2)$
Residual: $e_i = y_i - \hat{y_i}$
Residual Sum of Squares (RSS):
Variation because of factors other than the linear relationship between X and Y
$RSS = e_1^2 + e_2^2 + ... + e_n^2$
$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
$RSS = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$

Lesson 2: Linear Regression Continued

Least Squares Approach: Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes RSS
The minimizing values are:
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Standard Error: How estimator varies with repeated sampling.
Confidence Interval: 95% chance the interval will contain true value of β_1
$[\hat{\beta}_1 - t_{n-p-1,\alpha/2} * SE(\hat{\beta}_1), \hat{\beta}_1 + t_{n-p-1,\alpha/2} * SE(\hat{\beta}_1)]$
Hypothesis Testing: Is coef statistically significant
$H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$; $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$
p-value: If p-value is very small, the probability of seeing a t statistic more extreme than observed (assuming $\beta_1 = 0$) is very small.
Residual Standard Error: Estimate the variance of the noise. (How far response from regression line)
$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$
Total Sum of Squares (TSS):
$TSS = RegSS + RSS = \sum(y_i - \bar{y_i})^2$
Variation of the y_i from their mean $\bar{y}$
Regression Sum of Square (Reg SS): $RegSS = \sum(\hat{y_i} - \bar{y})^2$
Explained variation in linear relationship of X and Y.
$\mathbf{R^2}$: $R^2 = \frac{RegSS}{TSS} = \frac{TSS - RSS}{TSS}$
Coefficient Interpretation: A unit change in X_j is associated with a β_j change in Y on average, while all other variables stay fixed.
F Statistic: $F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)} = \frac{R^2/p}{1-R^2/(n-p-1)}; \sim F_{p,n-p-1}$
Is there a linear relationship between *all* of the X variables cosidered together and Y. H_0: All betas = 0 versus H_A: At least one $\beta_i \neq 0$
Qualitative Predictors: Create dummy variable for categorical features. $1 = female$ and $0 = male$
$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if ith person is female.} \\ \beta_0 + \epsilon_i, & \text{if ith person is male.} \end{cases}$

Lesson 3: Classification

Can't use linear regression for classification because it might produce probabilities less than zero or bigger than one, use logistic regression instead.
Logistic Regression: $p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$
Log Odds: $log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$
Likelihood Fucntion: $\prod_{i=1}^{n}[P(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i}$
We pick β_0 and β_1 to maximize the likelihood of the observed data.
NLL: $\sum_{i=1}^{n} y_i log(p(x_i)) + (1 - y_i) log(1 - p(x_i))$
Hypothesis Testing: $z = \frac{\beta_i - 0}{SE(\hat{\beta}_i)}$
Decision Boundary: The decision boundary between two classes is $p_{k=1}(x) = p_{k=2}(x)$ or $g_1(x) = g_2(x)$
Class Imbalance: Members of certain class(es) are rare. $< 5\%$ is severe, $< 20\%$ is marginal. Conditional imbalance when it is easy to predict Y from X is likely due to not enough samples.
Multinomial Regression: Logistic Regression with more than two classes
$Pr(Y = k|X) = \frac{e^{\beta_{0k} + \beta_{1k} X_1 + ... + \beta_{pk} X_p}}{\sum_{l=1}^{K} e^{\beta_{0l} + \beta_{1l} X_1 + ... + \beta_{pl} X_p}}$
Linear Discriminant Analysis: Does not suffer from instability when well-separated and handles $p > n$ well. Assume Gaussian distributions, and covariance is equivalent in all classes. Discriminant function is linear. Assign x to class with largest discriminant score.
Quadratic Discriminant Analysis: Covariance of each class is not assumed to be the same. Discriminant score function is quadratic.
Generic Discriminant Analysis:
$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$
$g_k(x) = log(P(Y = k|X = x))$ Denominators of all discriminant functions are the same, so if just assigning class, don't need to worry about denominator!

Lesson 3: Naive Bayes

Naive Bayes (classification) assumes features are independent in each class.

Bayes Rule:

$Pr(Y=k|X=x) = \frac{Pr(X=x|Y=k)Pr(Y=k)}{Pr(X=X)}$

$Pr(Y=k|X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$

For discrete features:

$f_k(x_i) = Pr(X_i = x_i|Y=k) = |x_{ik}|/N_k$

X_i is X_1 or X_2, while x_i is specific value in feature vector trying to predict.

For continuous features: Calculate sample mean and variance

$f_k(x_i) = Pr(X_i = x_i|Y=k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{(x_i-\mu_{ik})^2}{2\sigma_{ik}^2}}$

$f_k(x_1, x_2, ..., x_p) = \prod_{i=1}^{k} f_k(x_i)$

Classify sample to Y=k if $\pi_k \prod f_k(x_i)$ is maximum

Laplace Smoothing: $p(x_j|C_i) = \frac{N_{ji}+1}{N_i+c}$

Laplace smoothing if one of the conditional probabilities is zero

Lesson 4: Re-sampling

Cross validation: Split data into train and test. Split train into k folds. 1 of the k folds is used for validation (check test error of model), the rest of the folds are used to train the model.

LOOCV: $CV(n) = \frac{1}{n}\sum_{i=1}^{n} MSE_i$

CV: $CV(k) = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$

$LOOCV \downarrow$ bias $\uparrow$ variance across different samples of population

Stratified Cross Validation: Split data by target class, choose proportional from each for train, validation, test.

Bootstrap: Resample entire dataset b times with replacement.

Bootstrap Confidence Interval: Create B bootstrap datasets (each dataset size N, same as original). Calculate the mean of each b dataset. Order the sample means. The middle $(1-\alpha)B$ yield the $(1-\alpha)$ CI for the mean.

Lesson 5: Model Selection

Alternatives to Least Squares: Subset selection, Shrinkage, Dimension Reduction

Want to $\downarrow$ *variance* at the expense of bias!

Subset Selection: Pick subset of p predictors and fit model using least squared on reduced set of variables.

For k = 1,2,..p: Fit all p choose k models that contain exactly k predictions, pick the best model and call it M_k Select the best model $M_0 ... M_p$ using cross-validation predicted error.

Stepwise Selection: Forward selection and backward selection. Forward selection is the only viable subset method when p is very large. Backward requires $n > p$.

Estimate Test Error From Train Error: Does not require an estimate of the error variance $\sigma^2(RSE^2)$

($\downarrow C_p$, AIC, BIC) vs. $\uparrow$ Adjusted R^2

Mallow's $\mathbf{C_p}$ $= \frac{1}{n}(RSS + 2(p+1)RSE^2)$

AIC $= -2logL + 2(p+1)$

Use AIC for logistic regression (no RSS)

BIC $= \frac{1}{n}(RSS + log(n)(p+1)RSE^2)$

BIC places more penalty on models with many variables over Mallow C_p because $logn > 2$

Adjusted $\mathbf{R^2}$ = 1 - $\frac{RSS(n-d-1)}{TSS/(n-1)}$

Ridge Regression (l2): Shrinks coefficient estimates towards zero. Change the least squares objective, instead of minimizing RSS (in linear regression), minimize $RSS + \lambda\sum_{j=1}^{p} \beta_j^2$

$\lambda \to 0, \hat{\beta_{RR}} \to \hat{\beta_{LS}}; bias \downarrow\ variance \uparrow$

$\lambda \to \infty, \hat{\beta_{RR}} \to 0;\ bias \uparrow\ variance \downarrow$

Ridge Regression good at coefficient sharing, dealing with highly correlated features, bad at dealing with irrelevant. Includes all p predictors.

Lasso (l1): Shrinks coefficient estimates to zero, change LS objective to minimize $RSS + \lambda(|\beta_1| + |\beta_2| + ... + |\beta_p|)$

Elasticnet: combine L1 and L2 regularization, change LS objective to minimize $RSS + \lambda[\frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$

Dimension Reduction:

First PCA = $Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_1$.

To keep phi's from increasing: $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

Lesson 3: Metrics

False Positive Rate(Type I:) $FPR = \frac{FP}{FP+TN}$

False Negative Rate(Type II:) $FNR = \frac{FN}{FN+TP}$

Recall/Sensitivity/True Positive Rate: Rate the event of interest is predicted correctly for all samples having event. $Recall = \frac{TP}{TP+FN}$

Specificity/True Negative Rate: The rate that non-events are predicted correctly for all non-event samples. $Specificity = \frac{TN}{FP+TN}$

Precision: Ratio of true positives with respect to all detected positives. $Precision = \frac{TP}{TP+FP}$

Negative Predictive Value: Ratio of true negatives with respect to all detected negatives. $NPV = TN/(TN+FN)$

$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$

Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$

Multi-Class Metrics: Macro average across already calculated scores vs. pool all instances of class into one score for micro.

Misc

Changing the conditional probability threshold:

We can change the threshold from $Pr(Y=k|X) >= 0.5$ to some other value in $[0,1]$. By reducing the threshold we can reduce the False Negative Rate at the expense of False Positive Rate and overall error.

No Free Lunch: No single best optimization algorithm