# Statistical Learning Notes

Shehtab Zaman

April 23, 2020

## Contents

# 1 Regression

## 1.1 Ordinary Least Square Regression

- For a given input X and Response y, obtaining a matric $\beta$ such that the loss is minimized.
- The loss for OLS is:
$$L = \frac{1}{n}(y_i - f(x_i))^2 \tag{1}$$
- The residual sum of squares (RSS) is used to define the objective of the optimization problem
- The RSS is convex and therefor has a unique solution
$$RSS = (y - X\beta)^T(y - X\beta) \tag{2}$$
- The gradient equation for OLS is:
$$\nabla RSS(\beta) = 2 - X^T(y - X\beta) = 0 \tag{3}$$
- The solution the gradient equation is:
$$\hat{\beta} = (X^TX)^{-1}X^Ty \tag{4}$$
- OLS is an **unbiased** estimator
- OLS is essentially empirical risk minimization
- If the number of predictor is larger than observations, then overfitting starts to occur
- Highly correlated predictors will also provide incorrect solutions

## 1.2 Ridge Regression

- The $\beta$ for RR is obtained by minimizing the RSS with an additional *shrinkage penalty*
$$(y - X\beta)^T(y - X\beta) + \lambda||\beta^{(-0)}||_2 \tag{5}$$
$$||\beta^{(-0)}||_1 = \sum_{j=1}^{p} \beta_p^2 \tag{6}$$

- It is a constrained least square problem
- The interept term $\beta_0$ is never penalized
- The RR solution is **NOT** invariant to scaling of the input variables
- Typically you standardize the inputs so that they have same / similar scales
- RR also has a closed form solution for centered data:

$$\hat{\beta}_{\text{Ridge}} = (X^T X + \lambda I)^{-1} X^T y \tag{7}$$

- There exists $\lambda > 0$ such that the MSE of RR with that $\lambda$ is less than the OLS MSE

### 1.3 Lasso Regression

- Least Absolute Shrinkage and Selection Operator
- L2-norm (RR) does not do variable selection
- L0-Norm methods for variable selection are too computationally expensive as the objective is not convex
- Lasso is L1-norm Regression
- The Lasso objective function is:

$$(y - X\beta)^T (y - X\beta) + \lambda ||\beta^{(-0)}||_1 \tag{8}$$

$$||\beta^{(-0)}||_1 = \sum_{j=1}^{p} \beta_p \tag{9}$$

- The solution is not linear in y
- There is no closed-form solution
- Lasso results in coefficient estimates that are exactly 0
- Under orthonormal conditions on the data X, Lasso has a closed form solution
- Gradient Descent, Coordinate Descent can be used to obtain the solution for Lasso
- Variable selection consistency is not satisfied for lasso

### 1.3.1 Adaptive Lasso Regression

- Adaptive Lasso is *consistent* for more cases than Lasso
- Penalize more intensely those estimates which are more likely to be zero
- It is an approximatin to $l_q$ penaltiy where $q = 1 - \nu$ for $\nu \in (0.1)$

## 1.4 Elastic Net Regression

- Combination of Lasso and Ridge penalties
- The Rss is combined with the new penalty term

$$P_{\text{EN}}(\beta_j) = \frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \tag{10}$$

- It selects variables like lasso
- It shrinks together the coefficients of correlated predictors like ridge

## 1.5 Comparison of Models

The following table is a comparison of the linear regression and their various characteristics:

| Model | Closed-form | Penalty | Var. Sec. |
|---|---|---|---|
| OLS | T | N/A | F |
| RR | T | L2 | F |
| Lasso | F | L1 | T |
| Elastic | F | L1+L2 | T |

# 2 Classification

## 2.1 Logistic Regression

- Map the probabilities of the regression model with a logit function to obtain a minimization objective
- Each Generalied Linear Model (GLM) can be solved by minimizing a negative log-likelihood
- negative log-likelihood is similar to RSS in being analogous to empirical risk
- LR is a special case of a GLM
- Optimization methods used to solve LR:
  - Newton-Raphson Method
  - Gradient Descent
- NR update for LR:

  $$\beta^{(k+1)} \leftarrow \beta^k - [X^T W X]^{-1} X^T (y - p) \tag{11}$$

  The vector p = $(p(x_1), p(x_2)...p(x_n))^T$ is defined with the function,

  $$p(x) = \frac{e^{x^T \beta}}{1 + e^{x^T}\beta} \tag{12}$$

  and,

  $$\nabla p(x) = p(x)[1 - p(x)]x \tag{13}$$

  and the gradient:

  $$\nabla \ell(\beta) = X^T (y - p) \tag{14}$$

- The gradient update for LR:

  $$\beta^{(k+1)} \leftarrow \beta^k - \lambda X^T (y - p) \tag{15}$$

- Two encodings of labels
  - Deviance Loss: $y \in \{0, 1\}$
  - Logistic Loss: $y \in \{-1, 1\}$
- Mostly used as an inference tool. The goal is to understand the role of the input variables in explaining the outcome
- Good prediction performance for some datasets. Many other other methods which similar (or better) prediction performance

### 2.1.1 Multinomial Regression

- The goal is to model $P(Y = k|X)$ for k = 1,...,K classes.
- Choose one class refrence. Therefor, we get K-1 pariwise logisitc regression models
- There is also a symmetric alternative with K formulas

## 2.2 K-Nearest Neighbor Classifier

- Non parametric approach
- Very good for non-linear decision boundary and non Gaussian data
- Does not provide information on predictors
- Paramteric models, so there are assumptions on the distribution generating the data

## 2.3 Discriminant Analysis

- Assuming a Gaussian distribution on the data, LDA and QDA are used to classify using the Bayes Rule
- LDA can also be thought of geometrically. LDA projects the data such the class centroids of the projected data are spread out as much as possible
- Coefficients are significant

### 2.3.1 LDA

- Highly parametric method
- Assums Guassian with common covariance
- Simple linear boundary and posterior log-odd is a linear function
- Have issues with high-dimensional data

### 2.3.2 QDA

- Compromise between highly parametric LDA and non-parametric kNN. Assumes Gaussian nature of data
- The Covariances can be different in QDA
- QDA is more computationally difficult and a more complex model
- The decision boundary does not have to be linear

### 2.3.3 Naive Bayes Classifier

- Special case of LDA
- The assumption is the covariance matrix is diagonal

### 2.3.4 Nearest Centroid Classifier

- Special case of LDA
- The covariance matrix is replaced by the identity matrix

## 2.4 Support Vector Machine

### 2.4.1 Linear SVM

- Maximal margin hyperplanes are used as the basis for SVMs and aims to solve the drawbacks such that the sensitivity to a change in a single observation or non-slinearly seperable data
- Linear SVMs allow for some observations to be on the wrong side of the margin
- Define $\eta_i$ such that the 0ith observation is able to be on the wrong side of the margin and even the wrong side of the hyperplane
- SVM is Hinge Loss + $\ell_2$ penalty
- Hinge loss is a surrogate loss function, it is used for a convex approximation of the 0-1 loss
- SVMs perform better than LR for seperated classes
- When classes are overlapping, LR is better

- SVM cannot be used for inference
- SVM has extensions with $\ell_1$ norm (sparse) and 2-norm margin

### 2.4.2 Multiclass SVM

- There are a few approaches to multiclass SVMs
- One-Versus-One Classification
  - Construct K choose 2 binary SVMs, eachcompares a pair of classes.
  - Use a majority voting rule to select the class
- One-Versus-All Classification
  - Fit K Svms, each time comparing one of the K classes to the remaining K-1 classes
  - For a discriminant function $f_k(x)$, find the class k that maximizes $f_k(x)$
- Simultaneous Approach

### 2.4.3 Kernel SVM

- Extend SVMs to have non-linear separating boundaries
- Kernel Trick is a general mechanism for converting a linear classifier into a one that produces non-linear decision boundaries
- The idea is to map the original data to an new space, and generate linear boundaries in that space. The linear boundary in the new feature space is then mapped back to the original space, which is now non-linear.
- Kernel trick: replace the inner product of data observations with a generalization of the inner product or kernel
- No new calculations or alterations to the algorithm is needed. Just replace dot products with a new kernel
- Some Kernel Functions are:
  - linear kernel (linear SVM): $k(x, y) = x^T y$

  - polynomial kernel: $k(x, y) = (c + x^T y)^d$, where d is a positive integer
  - Radial Basis kernel (Gaussian): $k(x, y) = \exp -\gamma ||x - y||^2$
  - The feature space for a RBF kernel is infinite dimensional

### 2.5 Classification Measures

- In most cases we only care about the misclassification rate
- Bayes rule aims to minimize the misclassificaion rate
- Some values:
  - TN : True negative $y_p = y = 0$
  - TP: True positive $y_p = y = 1$
  - FN: False negative $y = 1, y_p = 0$
  - FP: False positive $y = 0, y_p = 1$
- Other measures are can also be useful (especially for binary classification):
  - False positive rate: $\frac{FP}{TN+FP}$
  - False negative rate: $\frac{FN}{TP+FN}$
  - Sensitivity/Recall/True Positive rate: $\frac{TP}{TP+FN}$
  - Specificity/ True Negative rate: $\frac{TN}{TN+FP}$
- ROC Curve: Displaying two types of errors for all possible thresholds
- Overall performance of the classifiers and all the thresholds are given by the Area under the ROC curve (AUC)
- Perfect classifier has an AUC of 1

### 2.6 Comparison of Models

# 3 Dimensionality Reduction

## 3.1 Principal Component Analysis

- Primarily used to find a low dimensional represnation that captures most of the variation in the data

- The new direction vector norms to 1

$$||v_j|| = 1 \tag{16}$$

- the direction vectors are orthognal so they form a new coordinate system

- The PC direction vector is called the loading vector

- The variance of the loading vector multiplied by the data is the variance explained by that PC

- The first explained by the i-th PC is always greater than or equal to the i+1th PC.

- The kth eigenvector is the same as kth PC direction vector

- The kth eigenvalue is the same as the variance explained by the kth PC

- Thus, PCA is a eigen-decomposition problem

- The eigen-decomposition of data X is equivalent to the singulr value decomposition (SVD) of the centered X

- Principal components provide low dimensional linear surfaces that are closest to the original observations

- PCA is unsupervised

- PC are linear combinations of p variables and are difficult to interpret

- scree plots and cumulative scree plots are used to determine how many components to keep when doing dimensionality reduction

- PC loading vectors are unique upto a sign flip

- PC scores are also unique, up to a sign flip

- PCA gives a global solution
- PCA is not invariant to scaling. Scaling changes PCA solution. Features with large scale contribute more to variance
- PCA can perform poorly when the number of features far exceeds the number of data points

### 3.1.1 Sparse PCA

- Goal of sparse PCA is zero out irrelevant features in the PC direction vectors
- Typically optimize PCA criterion with a sparcity encouraging penalty of V

### 3.1.2 Kernel PCA

- Replace the $XX^T$ in PCA by a Kernel Matrix K
- Only kernel PC scores available. No loadings

## 3.2 Independent Component Analysis

-

## 3.3 Non-negative Matrix Factorization

- The data, direction vectors and scores all non-negative
- The W and H matrices are often sparse
- Calculated by obtained non-negative least squares
- Only local solutions
- Factors and scores are not unique
- Can be considered a soft-clustering method. Entries in $w_i$ is the strength of memberships to different clusters
- Must supply the number of components beforehand

### 3.3.1 Archetypal Analysis

- Special Case for NMF
- In addition to the NMF requirements, there are two additional requirements:
    1. The rows of the W matrix sum to 1 .The data is within the convex hull of the archetypes or loading vectors *approximately*
    2. Archetypes themselves are convex combinations of the data points
- Related to k-means clustering
- K-means may be viewed as a special kind of NMF

## 3.4 Independent Component Analysis

- Decomposes the data matrix into components (factor) and lading vectors
- The Factors are modeled without assuming Gaussian random vectors
- The factors are aimed to be statistically independent

## 3.5 Multidimensional Scaling

- The goal of MDS is to visually represent proximities and distances between objects in a lower dimensional space
- The input is the matrix of similarities or dissimilarities (So a square nxn matrix)
- Distrance-preserving configurations are foind by optimizing a ”stress” function
- Can do non-linear dimension reduction
- Gradient descent used for finding solution
- Rotationally symmetric mappings
- Only relative distances matter
- Non-linear dimensional reduction
- Non-unique solution and sensitive to initial value

## 3.6 Strengths and Weaknesses

### 3.6.1 PCA

Strengths:

- Best linear dimension reduction
- Ordered / orthogonal components
- Unique, global solution

Weaknesses:

- Non-linear patterns. Uninterpretable
- Fails for very high dimensions where number of features far exceeds the number of observations

### 3.6.2 NMF

Strengths:

- The loadings are interpretable
- Linear dimension reduction
- Factors are not nested
- Archetypal Analysis and Pattern recognition very useful

Weaknesses:

- The factors are not ordered
- Loading vectors are not orthogonal
- Non-unique, non-global solution
- Must be run multiple times since the initialization matters

### 3.6.3 ICA

Strengths:

- OBtains independendent factors
- Specifically looks for non-Gaussian distributions

Weaknesses:

-

### 3.6.4 MDS

Strengths:

- Only requires matrix of similarity or dissimilarity
- Can be used with qualitative data (non-metric mapping)

Weaknesses:

- Sensitive to initial conditions
- Coordinates not meaningful

# 4 Unsupervised Learning

## 4.1 K-means Clustering

- One of the most popular iterative descent clustering
- Designed for the case that all variables are quantitative
- The squared Euclidean distance is used as the dissimilarity measure
- The Iterative Algorithm:
  - For a given clustering assignment, find the closest centroid for each observation and assign a new label to the observation
  - Given the new labels, calculate the new centroids of each cluster
- Repeat the algorithm until the labels don't change for the observations
- Always converges, possibly to a local solution
- Does not work well when cluster sizes are different
- Different initializations may result in different cluster assignments. Thus K-means usually requires to be run a multiple times with random initialization

## 4.2 Gaussian Mixture Model

- Model based clustering
- Data is a mixture of K subpopulation, each having its own distribution
- Each observation has a probability of arising from distribution j
- Gaussian Mixture Model: the probability of observation X being in cluster k is represented by a Gaussian
- Cluster assignment is the probability of the observation being in the cluster
- The Expectation Maximization algorithm is used to generate the clustering:
  - E-step: Compute the probability that the observations are in each cluster
  - M-step: Maximum likelihood estimation to estimate the mean and variance for ea h cluster

### 4.3 Hierarchical Clustering

- Gives clustering assignments for all K values. Each point gets assigned all K cluster values
- Nested Clusters
- Unique solution
- Agglomerative vs. Divisive fitting method
- Agglomerative
  - Starts with all points in their own clusters (n clusters)
  - Merge two clusters that have the smallest dissimilarity between them
-
  - Starts with all points in the same cluster
  - Split clusters that result in the biggest dissimilarity when split
- Interpreting a Dendogram
  - The leafs are the observations
  - similar lives share parent branches
  - The lower the split, the more similar the child leaves are
  - Height of the shared parent node indicates similarity between objects
- Linkage functions define the measure of dissimilarity between two clusters
- Some linkage functions:
  - Single Linkage: Minimum dissimilarity between points in two clusters
  - Complete Linkage: Maximum dissimilarity between points in two clusters
  - Average Linkage: Average dissimilarity between points in two clusters
  - ward's Linkage: Sum of squared distance to cluster center

### 4.4 Biclustering

- The objective is find meaningful groups of both observations and features
- Clustering both rows and columns of the data matrix
-

### 4.5 Convex Clustering

- Unique and global solution
- Fast algorithm
- Only one tuning parameter
- Family of clustering solutions
- Performance depen on weights
- Must have the entire data vector, The dissimilarity only is not enough

# 5 Optimization

## 5.1 Coordinate Descent

- CD takes each feature at a time and varies the coefficient such that the the temporary response is minimized with respect to the feature
- The process is repeated for each feature

## 5.2 Gradient Descent

- Gradient Descent is a general optimization strategy
- GD is a first order optimization method
- The objective is to solve the gradient equation:

$$\nabla\ell(\beta) = 0 \tag{17}$$

- GR updates using the gradient:

$$\beta^{(k+1)} \leftarrow \beta(k) - \gamma\nabla f(\beta^{(k}) \tag{18}$$

- GD directly updates towards the direction of steepest descent of f

## 5.3 Newton Raphson Method

- The Newton-Raphson method is motivated by Taylor expansion
- NR is a second-order optimization algorithm
- The objective is to solve the gradient equation:

$$\nabla\ell(\beta) = 0 \tag{19}$$

- The NR method updates the solution at each time step.

$$\beta^{(k+1)} \leftarrow \beta(k) - [H\ell(\beta(k))]^{-1}\nabla\ell(\beta(k)) \tag{20}$$

  where $H\ell$ is the Hessian matrix for $\ell$, such that

$$(H\ell)_{ij} = \frac{\partial^2\ell(\beta)}{\partial\beta_i\partial\beta_j} \tag{21}$$

- Requires fewer iterations when the Hessian is known and easy to compute

# 6 Model Assessment and Validation

## 6.1 Bias, Variance and Model Complexity

- The error in a model can be characterized as the squared bias plus the variance
- Linear models with ordinary least squares have zero bias but high variance
- Restricted fits such as Ridge and Lass have higher bias but the errors is balanced with lower variance
- $C_p$, AIC, and BIC
  - AIC: Akaike information criterion
  - The effect number of parameters is considered when minimizing model complexity
  - BIC: Bayesian information criterion
  - BIC tends to penalize complex models more haevily, giving preference to simpler models in selection
  - BIC is motivated by the Bayesian approach to model selection

## 6.2 Cross-Validation

- Usually a test case is not available as real world data is difficult to accumulate. We need a way to predict test error without relying solely on the training error rate.
- Important to note that the training error can often dramatically underestimate the testing error
- Cross-Validation is a method of estimating the test error rate by using a *hold out* subset of the data during the training process.

### 6.2.1 Validation Set

- The validation set approach estimates the test error by randomly dividing the available data into a *training set* and a *hold-out or validation set*
- The model is fit on the training set and the error on the validation set is used as a test set estimate

- Pros:
  - Conceptually easy
  - Easy to implement
- Cons:
  - The validation set error can vary highly depending on the observations included in the training set and validation set
  - A smaller portion of the data is used as training data and thus fewer observations are used to fit the model. This means validation set error tends to *overestimate* the test error for the model fit on the entire dataset.

### 6.2.2 Leave-One-Out Cross-Validation

- Leave-one-out cross-validation (LOOCV) is similar to the validation set approach while tackling the drawbacks of the validation set approach.
- The dataset is divided into two parts, with only one hold out observation.
- The model is fit on the rest of the dataset. The hold out observation is used to test the efficacy of the model.
- This process is repeated for the entire dataset, each time holding out a new observation.
- The final estimated error is the average of all the hold-out observation errors.
- Pros
  - Since all but one element of the dataset is used to fit the model, the bias is significantly lower than the validation set approach
  - This approach also tends to not overestimate the test error since most of the data is used for training
  - There is no random splitting of the data so the estimated error is always the same
- Cons

  - Very expensive to implement especially for large datasets. For a dataset with n number of observations, the model his to be fit n times, which may be computationally infeasible
  -

### 6.2.3 K-Fold Cross-Validation

-