ADVANCED STATISTICS FOR FINANCE

INDEX:

1. Linear Regression
2. KNN methods
3. Logistic Regression
4. Discriminant Analysis
5. Variable selection
6. Lasso and Ridge methods
7. Nonparametric Regression
8. Principal Component and Partial Least Squares Regression
9. Regression and Classification Trees
10. MonteCarlo methods
11. Likelihood Inference
12. Methods of estimation (MLE, Moments)
13. Introduction to Bayesian Statistics (Beta-Binomial example)

LINEAR REGRESSION

In statistical learning, we aim to predict an outcome **Y (dependent variable)** based on a vector of features X**T=(X1,X2,…,Xp) (independent variables). The outcome can be either quantitative** (e.g., stock price) or categorical (e.g., heart attack/no heart attack). We utilize a training set to observe outcomes and features, constructing a learner f(X) to predict the outcome y based on specific feature values.The scenario described is typical of supervised learning, where the outcome is measured in the training data. In contrast, unsupervised learning involves only features without measured outcomes, focusing on tasks like clustering and data simplification.In high-dimensional settings, such as when n=100 and p=2000, traditional statistical techniques may fail. For instance, the ordinary least squares (OLS) method becomes ineffective because the matrix XTX is not invertible, leading to infinitely many solutions. This necessitates modifications to statistical methods, including:

- Dimension Reduction: Reducing p to be smaller than n.
- Variable Selection: Removing variables with low correlation to the response.
- Shrinkage Methods: Techniques like ridge regression and lasso to constrain coefficients.

In supervised learning, variables can be quantitative or qualitative. The two primary tasks are:

- Regression: For quantitative responses.
- Classification: For categorical responses.

Common approaches include linear models using OLS or maximum likelihood estimation (MLE) and k-nearest neighbors (KNN). Each method has its strengths and weaknesses, which will be explored further.

Linear regression is a straightforward method where the model is expressed as:

$$Yi=\beta0+\beta1X1i+\beta2X2i+\ldots+\beta pXpi+\epsilon i$$

Properties of Linear Regression

- Interpretability: Easy to understand and explain.
- Computation: Straightforward to compute.
- Theoretical Properties: Well-established theoretical foundations.

Fitting the Model

The OLS method estimates coefficients by minimizing the residual sum of squares (RSS):

$$RSS(\beta)=\sum i=1\ n(yi-xiT\beta)2$$

The solution is given by:

$$\beta\hat{}=(XT)-1\ XT\ y$$

Linear regression is a fundamental statistical method used to model the relationship between a continuous outcome Y and one or more covariates (X1,X2,…,Xp). The model can be expressed as:

$$Y=f(X)+\epsilon=B0+B1X1+\ldots+BpXp+\epsilon$$

where $E(\epsilon)=0$ and $Var(\epsilon)=\sigma 2$. The Ordinary Least Squares (OLS) estimator $B=(X'X)-1X'y$ is known as the Best Linear Unbiased Estimator (BLUE) due to the Gauss-Markov theorem, which states that it is linear, unbiased, and has the minimum variance among all linear estimators.

The bias-variance trade-off is a crucial concept in model assessment. It highlights the balance between bias (error due to overly simplistic assumptions in the learning algorithm) and variance (error due to excessive complexity in the model). The total error can be decomposed into three components:

- Irreducible Error ($\sigma 2$)
- Bias ($E[f(X)-f\hat{}(X)]2$)
- Variance ($Var[f\hat{}(X)]$)

This trade-off is essential for understanding model performance and generalization.

Model assessment involves evaluating the performance of a selected model, while model selection focuses on choosing the best model for a specific task. Generalization is key; a model must perform well on unseen data, not just the training dataset.

Types of Errors

- Training Error: Measured through a loss function, such as:
    - Quadratic Loss: $L(Y,f(X))=(Y-f(X))2$
    - Absolute Loss: $L(Y,f(X))=|Y-f(X)|$
- Testing Error: The prediction error over an independent test set, which is crucial for assessing model performance.

Choosing the right subset of features is vital to avoid overfitting and to enhance model interpretability. Various methods exist for variable selection, including:

- Forward Selection: Starts with no predictors and adds them iteratively based on their contribution to reducing the residual sum of squares (RSS).

- Backward Elimination: Begins with all predictors and removes the least significant ones iteratively.

Classical Approaches

- Best Subset Selection: Evaluates all possible combinations of predictors, which can be computationally expensive.
- Stepwise Selection: A more efficient alternative that explores a restricted set of models.

Two primary approaches exist for estimating test error:

1. Indirect Estimation: Adjusting training error to account for overfitting.
2. Direct Estimation: Using validation sets or cross-validation techniques.

Several criteria help in selecting the best model:

- **Mallows' Cp**: Adjusts the RSS based on the number of parameters.
- **AIC (Akaike Information Criterion)**: Penalizes model complexity.
- **BIC (Bayesian Information Criterion)**: Similar to AIC but imposes a heavier penalty for larger models.
- **Adjusted R^2**: Accounts for the number of predictors, rewarding models that explain variance without unnecessary complexity.

Criterion	Formula
Mallows' Cp	$Cp = \frac{RSS + 2d\sigma^2}{n}$
AIC	$AIC = -2\log L(B) + 2d$
BIC	$BIC = -\frac{RSS + d\log(n)\sigma^2}{n}$
Adjusted R^2	$AdjustedR^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$

In conclusion, understanding linear regression, the bias-variance trade-off, model assessment, variable selection, and model selection criteria is essential for effective statistical modeling and prediction.

KNN METHODS

K-nearest neighbors (KNN) is a flexible method that predicts the value of Y based on the average of the k closest observations. The prediction is defined as:

$$\mathbf{Y\hat{}(x)=1k\sum j/xj\in Nk(x)yj}$$

Considerations for KNN

- Flexibility: KNN makes fewer assumptions about the functional form of the relationship.
- Choosing k: The choice of k significantly impacts the model's performance and can lead to overfitting if too small.

Statistical decision theory provides a framework for finding the optimal learner f(X) by minimizing a loss function L(Y,f(X)). Common loss functions include:

- Squared Error Loss: $L(Y,f(X))=(Y-f(X))2$
- Zero-One Loss: Used for classification tasks.

The expected prediction error (EPE) can be derived from the joint distribution p(X,Y):

$$\mathbf{EPE(f)=\iint L(y,f(x))p(x,y)\,dx\,dy}$$

In high-dimensional settings, the trade-off between bias and variance becomes critical. Linear models are stable but may introduce bias, while KNN is less biased but can suffer from instability in high dimensions, known as the curse of dimensionality.

The relationship between statistics and data science is complex. Some view data science as a rebranding of statistics, while others see it as a subset or a distinct field. Key differences include:

- Statistics: Focuses on understanding the data-generating process and making defensible analyses.
- Machine Learning: Primarily concerned with predictive modeling and performance on test sets, often disregarding model assumptions.

In conclusion, both fields share techniques but differ in goals and methodologies, emphasizing the importance of context in data analysis.

LOGISTIC REGRESSION

Classification involves predicting the category of a qualitative response variable Y based on a feature vector X. The goal is to build a function C(X) that predicts the value of Y from X. Often, it is more beneficial to estimate the probabilities of X belonging to each category rather than just providing a classification. For instance, estimating the probability of an insurance claim being fraudulent is more informative than a binary classification of fraudulent or not.

While linear regression can be used for binary outcomes by coding Y as 0 and 1, it may yield probabilities outside the [0, 1] range. Logistic regression is preferred as it models the probability of Y being 1 given X using the logistic function:

$$P(Y=1 \mid X=x)=e\beta 0+\sum j=1p\beta jxj1+e\beta 0+\sum j=1p\beta jxj$$

The logistic regression model is defined for binary outcomes, where the likelihood function for a sample of size n is given by:

$$L(\beta;y)= \prod i= 1\ nP(Yi=yi \mid Xi;\beta)$$

The maximum likelihood estimates (MLE) are obtained numerically, typically using software like R or Python.

Confidence Intervals and Hypothesis Testing

For large samples, the estimates of β are approximately normally distributed. Confidence intervals for βj can be calculated, and hypothesis tests can be performed to evaluate the significance of predictors.

-Cross-validation is employed to assess the predictive accuracy of the logistic regression model. Different models can be compared based on their cross-validated error rates.
The probit model is an alternative to logistic regression, using the cumulative distribution function (CDF) of a standard normal variable instead of the logistic function. It can be implemented in R using the `family=binomial(link="probit")` option.

-The probit model is an alternative to logistic regression, using the cumulative distribution function (CDF) of a standard normal variable instead of the logistic function. It can be implemented in R using the `family=binomial(link="probit")` option.

Linear regression is not suitable for binary outcomes as it can produce probabilities outside the range [0, 1]. Logistic regression is preferred as it models the probability of a binary outcome using the logistic function.

Statistical Theory of Logistic Regression

The logistic regression model is defined as: $E(Y|X=x)=P(Y=1|X=x)=e\beta 0+\sum j=1p\beta jxj1+e\beta 0+\sum j=1p\beta jxj$ where Y is binary, and β are the coefficients estimated through maximum likelihood estimation (MLE).

Likelihood Function

The likelihood function for logistic regression is given by: $L(\beta;y)=\prod i=1nP(Yi=yi;Xi=xi;\beta)=\prod i=1nT(xi,\beta)yi(1-T(xi,\beta))1-y$

Discriminant Analysis

When the number of categories of the response y is more than two, the logistic model is inappropriate. In these cases, we can use:

- **multinomial Logistic (Probit) either ordered or not ordered**
- **Discrimination Analysis**

If the distribution of the features (covariates) is approximately Gaussian, then Discriminant Analysis is better.

Discriminant Analysis (DA) will classify observations to the category with the highest posterior probability. DA provides more stable result than the Logistic model when classes are well separated for small values of the sample size n and features X are approximately Gaussian when there are more than two classes.

We know that, for $k = 1, 2$,

$$f_k(x) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

Assume for the moment that $\sigma_1 = \sigma_2 = \sigma$. Then

$$P(Y=k|X=x) = \frac{\pi_k \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2\right)}{\pi_1 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2\right) + \pi_2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)}$$

We will choose the class with the highest $P(Y = k|X = x)$.
Since the denominator is the same for each k, it is enough to compare the numerators.

For all k, the log-numerator is

$$\begin{aligned}
&\log \pi_k - \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}(x-\mu_k)^2 \\
&= \log \pi_k - \frac{1}{2\sigma^2}(x-\mu_k)^2 \\
&= \log \pi_k - \frac{1}{2\sigma^2}(x^2 - 2x\mu_k + \mu_k^2) \\
&= \log \pi_k - \frac{\mu_k^2}{2\sigma^2} + \frac{x\mu_k}{\sigma^2} = \delta_k(x)
\end{aligned}$$

- Note that $\delta_k(x)$ is a linear function of x.
- So we talk of Linear DA.
  It comes from the assumption of a unique value of σ (omoskedasticity!)

Do we choose $k = 1$ or $k = 2$?

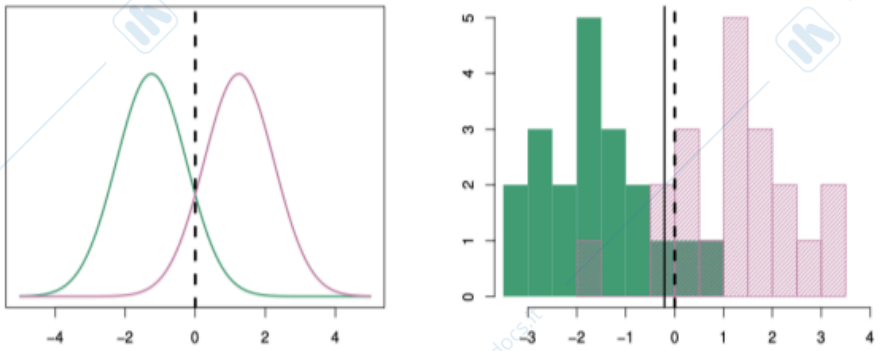
$$\delta_1(x) > \delta_2(x) \Leftrightarrow x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log\pi_1 - \left(x\frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log\pi_2\right)$$

$$\Leftrightarrow x\frac{\mu_1 - \mu_2}{\sigma^2} - \frac{\mu_1^2 - \mu_2^2}{2\sigma^2} > 0$$

$$\Leftrightarrow x\frac{\mu_1 - \mu_2}{\sigma^2} - \frac{(\mu_1 - \mu_2)(\mu_1 + \mu_2)}{2\sigma^2} > 0$$

$$\Leftrightarrow x > \frac{\mu_1 + \mu_2}{2}$$

An example



Example with $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$, and $\sigma^2 = 1$.

Typically we don't know these parameters; we just have the training data. In that case we simply estimate the parameters and plug them into the rule.

Variable selection

In high-dimensional regression, when the number of features p exceeds the sample size n, the least squares estimator becomes inaccurate and ill-defined due to the non-invertibility of XTX. A common solution is to perform variable selection, which involves including only a subset of features in the model. While this reduces variance, it may introduce bias by omitting important features.

Pre-processing the data by standardizing the variables is a common practice. This involves modifying each feature to have a mean of 0 and a variance of 1, represented mathematically as:

Xij=Xij−Xj/ Sj

This standardization simplifies the methods used in regression analysis.

THE RIDGE REGRESSION

Ridge regression is a technique that addresses the challenges of high-dimensional regression by introducing a bias-variance tradeoff. The least squares estimator is defined as the minimizer of the training error:

$$R\hat{}(\beta)=1\ n\sum i=1\ n(Yi-f(Xi,\beta))2=1n||Y-XB||2$$

In ridge regression, a new estimator is defined to minimize the penalized training error:

$$R2(\beta)=1\ n\sum i=1\ n(Yi-f(Xi,\beta))2+\lambda||\beta||2$$

where $\lambda>0$ is a tuning parameter. As λ increases, the coefficients β are shrunk towards zero, which helps in reducing variance.

The ridge estimator can be derived by setting the derivative of the penalized training error to zero:

$$\beta\hat{}(\lambda)=(XTX+\lambda Ip)-1XY$$

PREDICTION AND MODEL SELECTION

The tuning parameter λ creates a bias-variance tradeoff: larger values lead to low variance but high bias, while smaller values result in high variance and low bias. Cross-validation methods, such as K-fold or leave-one-out, are employed to select the optimal λ.

The prediction error can be computed as:

$$R\hat{}(\lambda)=1\ n\sum i=1\ n(Yi-f-i(Xi))2$$

where $f-i(Xi)$ is the ridge estimator computed after omitting the i-th observation.

LASSO REGRESSION

Lasso regression, introduced by Tibshirani in 1996, has largely replaced ridge regression due to its ability to perform variable selection automatically. The Lasso estimator is defined as:

$$R2(\beta)=1\ n\sum i=1\ n(Yi-f(Xi,\beta))2+\lambda\sum j=1p|\beta j|$$

This formulation encourages sparsity in the coefficient estimates, meaning many coefficients are driven to zero, simplifying the mod**el.**

Computational Aspects

The Lasso estimator does not have a closed-form solution, but efficient algorithms exist for its computation due to its convex optimization nature. The regularization parameter λ is typically chosen using K-fold cross-validation.

Comparison of Ridge and Lasso

While both ridge regression and Lasso have their advantages, they do not universally dominate each other. The Lasso is expected to perform better when the response is influenced by a small number of predictors. Cross-validation is essential for determining which method is more suitable for a given dataset.

BAYESIAN INTERPRETATION

From a Bayesian perspective, ridge and Lasso regression can be interpreted through prior distributions on the coefficient vector β. For ridge regression, assuming a Gaussian prior leads to minimizing the ridge equation, while for Lasso, a double exponential prior results in minimizing the Lasso equation.

Impact of Different Priors

The choice of prior can significantly affect the estimates obtained from ridge and Lasso regression. The graphical comparison of different priors illustrates their impact on the resulting coefficient estimates.

Conclusion

In high-dimensional regression, both ridge and Lasso regression provide valuable tools for managing the challenges posed by large feature sets. The choice between them depends on the specific characteristics of the data and the goals of the analysis.