Nonparametric Statistics

Daeyoung Lim[*]
Department of Statistics
Korea University

March 9, 2016

1 Regression analysis

- It is a method for investigating functional relationships among variables.
- (Ex1.) Whether the sale price of a home is related to physical characteristics of the building and taxes paid on the building.
- (Ex2.) Whether cigarette consumption is related to socioeconomic an demographic variables (such as ag, sex, education, income and price of cigarette).
- The relationship is expressed in the form of an equation connecting:

  $$\text{response variable} \leftarrow \text{predictor variables.}$$
- response variable = dependent variable, output
- predictor variables = covariates, regressors, factors, carriers, input etc
- Y: response variable
- $X_1, X_2, \ldots, X_p$: predictor variables
- The relationship between Y and $X_1, X_2, \ldots, X_p$ can be approximated by the regression model

  $$Y = f\left(X_1, X_2, \ldots, X_p\right) + \epsilon,$$

  where ϵ is a random error.
- The function $f\left(X_1, X_2, \ldots, X_p\right)$ describes the relationship between Y and $X_1, X_2, \ldots, X_p$.
- In essence, statistical modeling(or learning) refers to a set of approaches for estimating f.
- (**Parametric models**) An example is the linear regression model. Interpretable but less flexible. More appropriate for inference.
- (**Nonparametric models**)
  - Does not make explicit assumptions about the functional form of f.
  - Very flexible but less interpretable. more appropriate for prediction.
  - May require a very large number of observations to obtain an accurate estimate.
  - If the sample size is small, then parametric models are recommended.

[*] Prof. Sangbum Choi

- The simple and convenient approach is to consider a linear model. However, there is no general reason to think linear approximations ought to be good.

- If X takes on only a finite set of values, one can use

$$\hat{f}(x) = \frac{1}{\#\{i : x_i = x\}} \sum_{i:x_i=x} y_i.$$

- Unfortunately, this only works if X has a finite set of values. If X is continuous, the function will always be undersampled.

- k-nearest neighbor(KNN) fit for $\widehat{Y}$:

$$\widehat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

  where $N_k(x)$ is the neighbourhood of x dened by the k closest points x_i. If k is small, the estimated regression line will be wriggly, statistically speaking, '*overfitted*'. On the other hand, if k is large, then the function will be too smooth, statistically, '*underfitted*'. The main focus when using KNN regression is to choose the optimal value for k, which can be obtained in a data-driven fashion. In nonparametric statistics, k is called the '*tuning parameter*'. Other nonparametric models will also have some kind of tuning parameter(s).

- To use a KNN regression, we need to pick k somehow. This means we need to decide the degree of smoothing.

- As we increase k, we get smoother functions; in the limit $k = n$ and we just get back to constant.

- Thus, many nonparametric methods involve smoothing techniques.

- There will always exist a trade-off between flexibility and interpretability where one should sacrifice either one to gain more of the other.

- LASSO, which is used to select a meaningful subset of variables, is very much inclined toward interpretability. On the diametrical opposite, support vector machine or bagging drop most of the requirement for interpretability but attempt to achieve a high degree of flexibility.

1.1 Mean squared error or risk

- Suppose Y is a random variable and we try to predict Y by guessing a single value.

- What is the best guess? More formally, what is the optimal point forecast for Y?

- A reasonable starting point is to consider the (expected) mean squared error (MSE):

$$\text{Risk} = \text{MSE}(a) = \mathbb{E}\left[(Y-a)^2\right].$$

- (*Bias-Variance Trade-off*)

$$\begin{aligned} \text{MSE}(a) = \mathbb{E}\left[(Y-a)^2\right] &= [\mathbb{E}(Y-a)]^2 + \text{Var}(Y) \\ &= [\mathbb{E}Y - a]^2 + \text{Var}(Y) \\ \text{Risk(MSE)} &= \text{Bias}^2 + \text{Variance} \end{aligned}$$

- Now imagine we have two random variables (X, Y).

- We may want our prediction to be a function $f(X)$. Consider

$$\begin{aligned}\text{MSE}[f(x)] &=\equiv \mathbb{E}\left[(Y-f(x))^2\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[(Y-f(X))^2\middle|X\right]\right]\\ &= \mathbb{E}\left[\text{Var}(Y|X)+\mathbb{E}\left[Y-f(X)|X\right]\right]\end{aligned}$$

- **Inference** procedures concern constructing the estimate $\hat{f}$ for f, using a set of data sets:

$$\{(x_1,y_1),\ldots,(x_n,y_n)\},$$

  which is often called as training data in statistical learning.

- **Prediction** procedures make a prediction with $\hat{f}(x_0)$ for y_0, where (x_0,y_0) is a new observation. Note that (x_0,y_0) has no contribution to estimating $\hat{f}$.
- In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained.
- In this setting, we can predict $Y=f(X)+\epsilon$ using $\widehat{Y}=\hat{f}(X)$ where $\hat{f}$ represents our estimate for f.
- The accuracy of $\widehat{Y}$ as a prediction for Y depends on two quantities: reducible error and irreducible error.
- The best prediction can minimize the MSE:

$$\begin{aligned}\mathbb{E}\left[\left(Y-\widehat{Y}\right)^2\right] &= \mathbb{E}\left[f(X)+\epsilon-\hat{f}(X)\right]^2\\ &= \underbrace{\left[f(X)-\hat{f}(X)\right]^2}_{\text{reducible error}}+\underbrace{\text{Var}(\epsilon)}_{\text{irreducible error}}\end{aligned}$$

After fitting the model with the training dataset, we can evaluate how well the fitted model works with newly obtained test data. In a very simplistic manner, inference refers to the process where we use the training data to fit the model whereas prediction refers to using new data to get the predicted values.

1.2 Supervised vs Unsupervised learning

- Most statistical learning problems fall into one of two categories: supervised or unsupervised.
- **Supervised** learning: For predictor measurement(s) $x_i, i=i,\ldots,n$, there is an associated response y_i.
- **Unsupervised** learning: For every observation $i=1,\ldots,n$, we observe a vector of measurements x_i but no associated response y_i.
  - It includes clustering analysis, which is to ascertain whether the observations fall into relatively distinct groups.

1.3 Regression vs. Classification

- Variables can be characterized as either quantitative or qualitative:
  - **Quantitative**: height, income, age, stock price, etc
  - **Qualitative**: gender (male, female), marital status (single, married, or divorced)
- We tend to refer to problems with a quantitative response as regression problems, while those involving a qualitative response are often referred to as classification problems.

1.4 Measuring the quality of fit

- In order to evaluate the performance of a statistical method, we need some way to measure how well its predictions actually match the observed data.
- To this aim, one might consider the MSE.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}\left(x_i\right)\right)^2$$

- Note that estimation of $\hat{f}$ is based on the training data. But in general, we do not really care how well the method works on the training data.
- Rather, we are interested in the accuracy of the predictions when applying the method to previously unseen test data.
- We wish our estimate $\hat{f}$ has a good predictive power:

$$y_0 \approx \hat{f}\left(x_0\right),$$

  where (x_0, y_0) is a new observation.

- If we had a large number of test observations, then we could compute the so-called test MSE:

$$\text{Ave}\left\{\hat{f}\left(x_0\right) - y_0\right\}^2 = \frac{1}{m}\sum_{i=1}^{m}\left\{\hat{f}\left(x_0^i - y_0^i\right)\right\}^2,$$

  which is the average squared prediction error for these test observations.

- We'd like to select the model for which **the test MSE is as small as possible**.
- We should select a model that minimizes test mse is as small as possible.
- Instead, can we simply select a statistical method that minimizes the training MSE? DOes it also minimize the test MSE?
- Unfortunately, there is a fundamental problem with this strategy. (The answer is NO!)
- For example, suppose that data were generated from

$$\text{true} : Y = 10 + 2x + x^2 + \epsilon$$

- Then we applied the following models to the data:
  - (a) $Y = \beta_0 + \beta_1 x + \epsilon$
  - (b) $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
  - (c) $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$
  - (d) $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \epsilon$
- Obviously, model (b) should be selected, because it includes the case of the true model.
- You can minimize the training MSE as much as possible by using more flexible models (i.e., letting $d \to \infty$).
- This is not a sensible model selection approach.
- Instead, we **should** use the test MSE to measure the performance of the method.
- As model flexibility increases, the training MSE will decrease but the test MSE may not.
- Too flexible a model will result in overfitting. Too restrictive a model will result in underfitting.

2 Bias-Variance trade-off

- The (expected test) MSE can be decomposed as, for a given x_0,

  $$\mathbb{E}\left(y_0 - \hat{f}\left(x_0\right)\right)^2 = \text{Var}\left(\hat{f}\left(x_0\right)\right) + \left[\text{Bias}\left(\hat{f}\left(x_0\right)\right)\right]^2 + \text{Var}\left(\epsilon\right)$$

  where the bias is the difference between the fitted value and the true value for each data point. It is well-observed that a model with high bias will have low variance and vice versa.

- To understand the bias-variance trade-off better, let f be a pdf and consider estimating $f(0)$. Let $h > 0$ be a small number.

- Then we can show that

  $$\text{Bias} \approx \frac{f''\left(0\right)h^2}{24}, \text{Variance} \approx \frac{f\left(0\right)}{nh}.$$

- Therefore,

  $$\text{MSE} = \text{Bias}^2 + \text{Variance} \approx \frac{\left(f''\left(0\right)\right)^2 h^4}{576} + \frac{f\left(0\right)}{nh} \equiv Ah^4 + \frac{B}{nh}.$$

  To prove this,

  $$\begin{aligned} \mathbb{P}_h \equiv \mathbb{P}\left(-\frac{h}{2} < x < \frac{h}{2}\right) &= \int_{-h/2}^{h/2} f\left(x\right)\, dx \approx f\left(0\right)h \\ f\left(0\right) &\approx \frac{\mathbb{P}_h}{h} \\ X = \#\text{ of observations in }\left(-\frac{h}{2}, \frac{h}{2}\right) &\sim \text{Bin}\left(n, \mathbb{P}_h\right) \\ \mathbb{E}\left[X\right] &= n\mathbb{P}_h, \text{Var}\left[X\right] = n\mathbb{P}_h\left(1 - \mathbb{P}_h\right) \\ \hat{f}\left(0\right) &\approx \frac{\widehat{\mathbb{P}_h}}{h} = \frac{X}{nh} \end{aligned}$$

By Taylor expansion,

$$
\begin{aligned}
f\left(X\right) &\approx f\left(0\right) + Xf'\left(0\right) + \frac{X^2}{2}f''\left(0\right) \\
\mathbb{P}_h &= \int_{-h/2}^{h/2} f\left(x\right)\,dx \approx \int_{-h/2}^{h/2}\left(f\left(0\right) + xf'\left(0\right) + \frac{x^2}{2}f''\left(0\right)\right)dx \\
&\approx hf\left(0\right) + \frac{h^3}{24}f''\left(0\right) \\
\mathbb{E}\left[\hat{f}\left(0\right)\right] &\approx \frac{\mathbb{E}\left[X\right]}{nh} = \frac{\mathbb{P}_h}{h} \approx f\left(0\right) + \frac{h^2}{24}f''\left(0\right) \\
\text{Bias} &= \mathbb{E}\left[\hat{f}\left(0\right)\right] - f\left(0\right) \approx \frac{h^2}{24}f''\left(0\right) \\
\text{Var}\left[\hat{f}\left(0\right)\right] &\approx \frac{\text{Var}\left[X\right]}{n^2h^2} = \frac{n\left(1-\mathbb{P}_h\right)\mathbb{P}_h}{n^2h^2} \\
&\approx \frac{\mathbb{P}_h}{nh^2} \quad \left(\because \mathbb{P}_1 \approx 0,\ 1-\mathbb{P}_h \approx 1\right) \\
&\approx \frac{hf\left(0\right) + \frac{h^3}{24}f''\left(0\right)}{nh^2} \\
&= \frac{f\left(0\right)}{nh} + \frac{f''\left(0\right)h}{24n} \\
&\approx \frac{f\left(0\right)}{nh} \\
\text{MSE} &= \text{Bias}^2 + \text{Variance} \\
&= \frac{h^4}{24}\left(f''\left(0\right)\right)^2 + \frac{f\left(0\right)}{nh} \\
&\equiv Ah^4 + \frac{B}{nh}
\end{aligned}
$$

3 Classification Problems

- A severely injured patient is admitted to a trauma center. Should treat massibe blood transfusion or not?

$$
y_i = \begin{cases} 1, & \text{massive transfusion} \\ 0, & \text{no massive transfusion} \end{cases}
$$

- (*Training error rate*)

$$
\text{Training error rate} = \frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \hat{y}_i\right)
$$

- (*Testing error rate*)

$$
\text{Testing error rate} = \text{Ave}\left(I\left(y_i \neq \hat{y}_i\right)\right)
$$

- (*Bayes classifier*) Bayes classifier assign a subject with x_0 to the class j, for which $\mathbb{P}\left(Y = j | X = x_0\right)$ is the largest.
- In theory, the Bayes classifier is optimal.
- The Bayes classifier produces the lowest possible test error rate, called the *Bayes error rate.*
- In general, the overall Bayes error rate is given by

$$
1 - \mathbb{E}\left(\max_j \mathbb{P}\left(Y = j | X = x_0\right)\right)
$$

- However, it depends on unknown conditional probability $\mathbb{P}\left(Y = j | X = x_0\right)$, so computing Bayes classifier is impossible for real data. One alternative to Bayes classifier would be again KNN classifier.
- (KNN) Choosing too small a number for k, the model will end up overfitting the data. This may yield small bias but recall the bias-variance trade-off. If overfitting occurs, the model will most likely not be able to capture enough variability thereby firing errors once new data come in. On the other hand, if too large a number for k is chosen, the model will wind up underfitting the data, only to find that it does not give us satisfactory performance or accuracy.