Lecture Notes 15
Prediction

Chapters 13, 22, 20.4.

1 Introduction

Prediction is covered in detail in 36-707, 36-701, 36-715, 10/36-702. Here, we will just give an introduction.

We observe *training data* $Z_1, \ldots, Z_n \sim P$ where $Z_i = (X_i, Y_i)$ where $X_i \in \mathbb{R}^d$. Given a new pair $Z = (X, Y)$ we want to predict Y from X. There are two common versions:

1. $Y \in \{0, 1\}$. This is called *classification*, or *discrimination*, or *pattern recognition*. (More generally, Y can be discrete.)

2. $Y \in \mathbb{R}$. This is called *regression*.

For classification we will use the following loss function. Let $h(x)$ be or prediction of Y when $X = x$. Thus $h(x) \in \{0, 1\}$. The function h is called a **classifier**. The classification loss is $I(Y \neq h(X))$ and the the **classification risk** is

$$R(h) = \mathbb{P}(Y \neq h(X)) = \mathbb{E}(I(Y \neq h(X))).$$

For regression, suppose our prediction of Y when $X = x$ is $g(x)$. We will use the squared error prediction loss $(Y - g(X))^2$ and the risk is

$$R(g) = \mathbb{E}(Y - g(X))^2.$$

Notation: We write $X_i = (X_i(1), \ldots, X_i(d))$. Hence, $X_i(j)$ is the j^{th} feature for the i^{th} observation.

2 The Optimal Regression Function

Suppose for the moment that we know the joint distribution $p(x, y)$. Then we can find the best regression function.

Theorem 1 *$R(g)$ is minimized by*

$$m(x) = \mathbb{E}(Y|X = x) = \int y\, p(y|x) dy.$$

Proof. Let $g(x)$ be any function of x. Then

$$\begin{aligned}
R(g) &= \mathbb{E}(Y-g(X))^2 = \mathbb{E}(Y-m(X)+m(X)-g(X))^2 \\
&= \mathbb{E}(Y-m(X))^2 + \mathbb{E}(m(X)-g(X))^2 + 2\mathbb{E}((Y-m(X))(m(X)-g(X))) \\
&\geq \mathbb{E}(Y-m(X))^2 + 2\mathbb{E}((Y-m(X))(m(X)-g(X))) \\
&= \mathbb{E}(Y-m(X))^2 + 2\mathbb{E}\mathbb{E}\left((Y-m(X))(m(X)-g(X)) \,\middle|\, X\right) \\
&= \mathbb{E}(Y-m(X))^2 + 2\mathbb{E}\Big((\mathbb{E}(Y|X)-m(X))(m(X)-g(X))\Big) \\
&= \mathbb{E}(Y-m(X))^2 + 2\mathbb{E}\Big((m(X)-m(X))(m(X)-g(X))\Big) \\
&= \mathbb{E}(Y-m(X))^2 = R(m).
\end{aligned}$$

■

Of course, we do not know $m(x)$ so we need to find a way to predict Y based on the training data.

3 Linear Regression

The simplest approach is to use a parametric model. In particular, the *linear regression model* assumes that $m(x)$ is a linear function of $x = (x(1), \ldots, x(d))$. That is, we use a predictor of the form $m(x) = \beta_0 + \sum_j \beta(j)x(j)$. If we define $x(1) = 1$ then we can write this more simply as $m(x) = \beta^T x$. In what follows, I always assume that the intercept has been absorbed this way.

3.1 A Bad Approach: Assume the True Model is Linear

On approach is to assume that the the true regression function $m(x)$ is linear. Hence, $m(x) = \beta^T x$ and we can then write

$$Y_i = \beta^T X_i + \epsilon_i$$

where $\mathbb{E}[\epsilon_i] = 0$. This model is certainly wrong, so let's proceed with caution.

The least squares estimator $\widehat{\beta}$ is defined to be the β that minimizes

$$\sum_{i=1}^{n} (Y_i - X_i^T\beta)^2.$$

Theorem 2 *Let $\mathbb{X}$ be the $n \times d$ matrix with $\mathbb{X}(i,j) = X_i(j)$ and let $\mathbb{Y} = (Y_1, \ldots, Y_n)$. Suppose that $\mathbb{X}^T\mathbb{X}$ is invertible. Then the least squares estimator is*

$$\widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}.$$

Theorem 3 *Suppose that the linear model is correct. Also, suppose that* $\text{Var}(\epsilon_i) = \sigma^2$ *and that* X_i *is fixed. Then* $\widehat{\beta}$ *is unbiased and has covariance* $\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$. *Under some regularity conditions,* $\widehat{\beta}$ *is asymptotically Normally distributed. If the* $\epsilon_i's$ *are* $N(0, \sigma^2)$ *then,* $\widehat{\beta}$ *has a Normal distribution.*

Continuing with the assumption that the linear model is correct, we can also say the following. A consistent estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{\text{RSS}}{n-p}$$

and

$$\frac{\sqrt{n}(\widehat{\beta}_j - \beta_j)}{s_j} \rightsquigarrow N(0,1)$$

where the standard error s_j is the j^{th} diagonal element of $\widehat{\sigma}^2\mathbb{X}^T\mathbb{X}$. To test $H_0 : \widehat{\beta}_j = 0$ versus $H_1 : \widehat{\beta}_j \neq 0$ we reject if $|\widehat{\beta}_j|/s_j > z_{\alpha/2}$. An approximate $1-\alpha$ confidence interval for β_j is

$$\widehat{\beta}_j \pm z_{\alpha/2}s_j.$$

Theorem 4 *Suppose that the linear model is correct and that* $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$. *Then the least squares estimator is the maximum likelihood estimator.*

3.2 A Better Approach: Assume the True Model is Not Linear

Now we switch to more reasonable assumptions. We assume that the linear model is wrong and that X is random. The least squares estimator still has good properties. Let β_* minimize

$$R(\beta) = \mathbb{E}(Y - X^T\beta)^2.$$

We call $\ell_*(x) = x^T\beta_*$ the *best linear predictor*. It is also called the *projection parameter*.

Lemma 5 *The value of* β *that minimizes* $R(\beta)$ *is*

$$\beta = \Lambda^{-1}\alpha$$

where $\Lambda = \mathbb{E}[X_iX_i^T]$ *is a* $d \times d$ *matrix, and* $\alpha = (\alpha(1), \ldots, \alpha(d)$ *where* $\alpha(j) = \mathbb{E}[Y_iX_i(j)]$.

The plug-in estimator β is the least squares estimator

$$\widehat{\beta} = \widehat{\Lambda}^{-1}\widehat{\alpha} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

where

$$\widehat{\Lambda} = \frac{1}{n}\sum_i X_iX_i^T, \quad \widehat{\alpha} = \frac{1}{n}\sum_i Y_iX_i.$$

In other words, the least-squares estimator is the plug-in estimator. We can write $\beta = g(\Lambda, \alpha)$. By the law of large numbers, $\widehat{\Lambda} \xrightarrow{P} \Lambda$ and $\widehat{\alpha} \xrightarrow{P} \alpha$. If Λ is invertible, then g is continous and so, by the continous mapping theorem,

$$\widehat{\beta} \xrightarrow{P} \beta.$$

(This all assumes d is fixed. If d increases with n then we need different theory that is discussed in 10/36-702 and 36-707.) By the delta-method,

$$\sqrt{n}(\widehat{\beta} - \beta) \rightsquigarrow N(0, \Gamma)$$

for some Γ. There is a convenient, consistent estimator of Γ, called the *sandwich estimator* given by

$$\widehat{\Gamma} = \widehat{\Lambda}^{-1} M \widehat{\Lambda}^{-1}$$

where

$$M = \frac{1}{n}\sum_{i=1}^{n} r_i^2 X_i X_i^T$$

where $r_i = Y_i - X_i^T\widehat{\beta}$. Hence, an asymptotic confidence interval for $\beta(j)$ is

$$\widehat{\beta}(j) \pm \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\widehat{\Gamma}(j,j)}.$$

Another way to construct a confidence set is to use the bootstrap. In particular, we use the pairs bootstrap which treats each pair (X_i, Y_i) as one observation. The confidence set is

$$C_n = \left\{ \beta : \ ||\beta - \widehat{\beta}||_\infty \leq \frac{t_\alpha}{\sqrt{n}} \right\}$$

where t_α is defined by

$$\mathbb{P}\big(\sqrt{n}||\widehat{\beta}^* - \widehat{\beta}||_\infty > t_\alpha \ \big| \ Z_1, \ldots, Z_n\big) = \alpha$$

where $Z_i = (X_i, Y_i)$. In practice, we approximate this with

$$\mathbb{P}\big(\sqrt{n}||\widehat{\beta}^* - \widehat{\beta}||_\infty > t_\alpha \ \big| \ Z_1, \ldots, Z_n\big) \approx \frac{1}{B}\sum_{j=1}^{B} I\big(\sqrt{n}||\widehat{\beta}_j^* - \widehat{\beta}||_\infty > t_\alpha\big).$$

There is another version of the bootstrap which bootstraps the residuals $\widehat{\epsilon}_i = Y_i - X_i^T\widehat{\beta}$. However, this version is only valid if the linear model is correct.

3.3 The Geometry of Least Squares

The *fitted values* or *predicted values* are $\widehat{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^T$ where

$$\widehat{Y}_i = X_i^T \widehat{\beta}.$$

Hence,

$$\widehat{Y} = \mathbb{X}\widehat{\beta} = HY$$

where

$$H = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$$

is called the *hat matrix*.

Theorem 6 *The matrix H is symmetric and idempotent: $H^2 = H$. Moreover, HY is the projection of Y onto the column space of $\mathbb{X}$.*

4 Nonparametric Regression

Suppose we want to estimate $m(x)$ where we only assume that m is a smooth function. The kernel regression estimator is

$$\widehat{m}(x) = \sum_i Y_i \, w_i(x)$$

where

$$w_i(x) = \frac{K\left(\frac{||x - X_i||}{h}\right)}{\sum_j K\left(\frac{||x - X_j||}{h}\right)}.$$

Here K is a kernel and h is a bandwidth. The properties are simialr to that of kernel density estimation. The properties of $\widehat{m}$ are similar to the kernel density estimator and are discussed in more detail in the 36-707 and in 10-702. An example is shown in Figure 1.

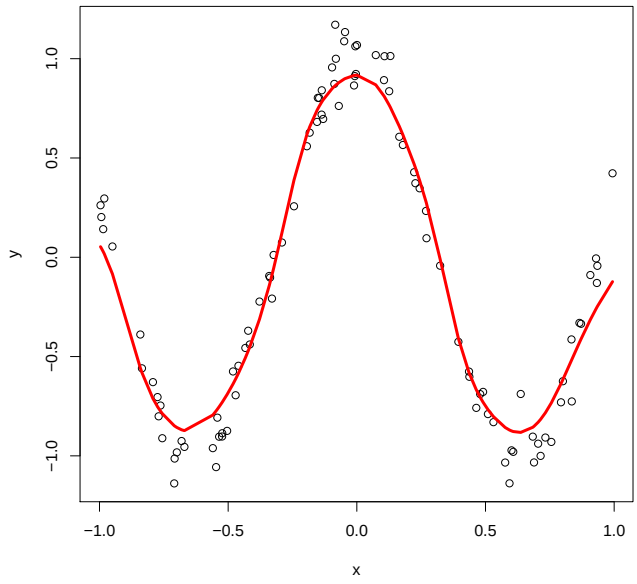

Figure 1: A kernel regression estimator.

5 Classification

The best classifier is the so-called *Bayes classifier* defined by:

$$h_B(x) = I(m(x) \geq 1/2)$$

where $m(x) = \mathbb{E}(Y|X = x)$. (This has nothing to do with Bayesian inference.)

Theorem 7 *For any h, $R(h) \geq R(h_B)$.*

Proof. For any h,

$$\begin{aligned} R(h) - R(h_B) &= \mathbb{P}(Y \neq h(X)) - \mathbb{P}(Y \neq h_B(X)) \\ &= \int \mathbb{P}(Y \neq h(x)|X = x)p(x)dx - \int \mathbb{P}(Y \neq h_B(x)|X = x)p(x)dx \\ &= \int \left(\mathbb{P}(Y \neq h(x)|X = x) - \mathbb{P}(Y \neq h_B(x)|X = x)\right) p(x)dx. \end{aligned}$$

We will show that

$$\mathbb{P}(Y \neq h(x)|X = x) - \mathbb{P}(Y \neq h_B(x)|X = x) \geq 0$$

for all x. Now

$$\begin{aligned}
\mathbb{P}(Y \neq h(x)|X=x) &- \mathbb{P}(Y \neq h_B(x)|X=x) \\
&= \left(h(x)\mathbb{P}(Y \neq 1|X=x) + (1-h(x))\mathbb{P}(Y \neq 0|X=x) \right) \\
&\quad - \left(h_B(x)\mathbb{P}(Y \neq 1|X=x) + (1-h_B(x))\mathbb{P}(Y \neq 0|X=x) \right) \\
&= (h(x)(1-m(x)) + (1-h(x))m(x)) \\
&\quad - (h_B(x)(1-m(x)) + (1-h_B(x))m(x)) \\
&= 2(m(x)-1/2)(h_B(x)-h(x)) \geq 0
\end{aligned}$$

since $h_B(x) = 1$ if and only if $m(x) \geq 1/2$. ■

The most direct approach to classification is *empirical risk minimization* (ERM). We start with a set of classifiers $\mathcal{H}$. Each $h \in \mathcal{H}$ is a function $h : x \to \{0, 1\}$. The *training error* or *empirical risk* is

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq h(X_i)).$$

We choose $\widehat{h}$ to minimize $\widehat{R}$:

$$\widehat{h} = \text{argmin}_{h \in \mathcal{H}} \widehat{R}(h).$$

For example, a linear classifier has the form $h_\beta(x) = I(\beta^T x \geq 0)$. The set of linear classifiers is $\mathcal{H} = \{h_\beta : \ \beta \in \mathbb{R}^p\}$.

Theorem 8 *Suppose that $\mathcal{H}$ has VC dimension $d < \infty$. Let $\widehat{h}$ be the empirical risk minimizer and let*

$$h_* = \text{argmin}_{h \in \mathcal{H}} R(h)$$

be the best classifier in $\mathcal{H}$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(R(\widehat{h}) > R(h_*) + 2\epsilon) \leq c_2 n^d e^{-nc_2\epsilon^2}$$

for some constnts c_1 and c_2.

Proof. Recall that

$$\mathbb{P}(\sup_{h \in \mathcal{H}} |\widehat{R}(h) - R(h)| > \epsilon) \leq c_2 n^d e^{-nc_2\epsilon^2}.$$

But when $\sup_{h \in \mathcal{H}} |\widehat{R}(h) - R(h)| \leq \epsilon$ we have

$$R(\widehat{h}) \ \leq \ \widehat{R}(\widehat{h}) + \epsilon \leq \widehat{R}(h_*) + \epsilon \leq R(h_*) + 2\epsilon. \quad \square$$

■

Empirical risk minimization is difficult because $\widehat{R}(h)$ is not a smooth function. Thus, we often use other approaches. One idea is to use a *surrogate loss function.* To expain this idea, it will be convenient to relabel the Y_i's as being +1 or -1. Many classifiers then take the form

$$h(x) = \text{sign}(f(x))$$

for some $f(x)$. For example, linear classifiers have $f(x) = x^T\beta$. Th classification loss is then

$$L(Y, f, X) = I(Yf(X) < 0)$$

since an error occurs if and only if Y and $f(X)$ have different signs. An example of surrogate loss is the hinge function

$$(1 - Yf(X))_+.$$

Instead of minimizing classification loss, we minimize

$$\sum_i (1 - Y_i f(X_i))_+.$$

The resulting classifier is called a *support vector machine.*

Another approach to classification is *plug-in clasification.* We replace the Bayes rule $h_B = I(m(x) \geq 1/2)$ with

$$\widehat{h}(x) = I(\widehat{m}(x) \geq 1/2)$$

where $\widehat{m}$ is an estimate of the regression function. The estimate $\widehat{m}$ can be parametric or nonparametric.

A common parametric estimator is *logistic regression.* Here, we assume that

$$m(x; \beta) = \frac{e^{x^T\beta}}{1 + e^{x^T\beta}}.$$

Since Y_i is Bernoulli, the likeihood is

$$L(\beta) = \prod_{i=1}^n m(X_i; \beta)^{Y_i}(1 - m(X_i; \beta))^{1-Y_i}.$$

We compute the mle $\widehat{\beta}$ numerically. See Section 12.3 of the text.

What is the relationship between classification and regression? Generally speaking, **classification is easier**. This follows from the next result.

Theorem 9 *Let $m(x) = \mathbb{E}(Y|X = x)$ and let $h_m(x) = I(m(x) \geq 1/2)$ be the Bayes rule. Let g be any function and let $h_g(x) = I(g(x) \geq 1/2)$. Then*

$$R(h_g) - R(h_m) \leq 2\sqrt{\int |g(x) - m(x)|^2 dP(x)}.$$

Proof. We showed earlier that

$$R(h_g) - R(h_m) = \int [\mathbb{P}(Y \neq h_g(x)|X = x) - \mathbb{P}(Y \neq h_m(x)|X = x)]\, dP(x)$$

and that

$$\mathbb{P}(Y \neq h_g(x)|X = x) - \mathbb{P}(Y \neq h_m(x)|X = x) = 2(m(x) - 1/2)(h_m(x) - h_g(x)).$$

Now

$$2(m(x) - 1/2)(h_m(x) - h_g(x)) = 2|m(x) - 1/2|\ I(h_m(x) \neq h_g(x)) \leq 2|m(x) - g(x)|$$

since $h_m(x) \neq h_g(x)$ implies that $|m(x) - 1/2| \leq |m(x) - g(x)|$. Hence,

$$\begin{aligned} R(h_g) - R(h_m) &= 2\int |m(x) - 1/2| I(h_m(x) \neq h_g(x)) dP(x) \\ &\leq 2\int |m(x) - g(x)| dP(x) \\ &\leq 2\sqrt{\int |g(x) - m(x)|^2 dP(x)} \end{aligned}$$

where the last setp follows from the Cauchy-Schwartz inequality. □ ■

Hence, if we have an estimator $\widehat{m}$ such that $\int |\widehat{m}(x) - m(x)|^2 dP(x)$ is small, then the excess classification risk is also small. But the reverse is not true.