Lecture Notes 23
Classification

Suppose we observe $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ where $X_i \in \mathbb{R}^d$ and $Y_i \in \{1, \ldots, k\}$. Let $(X, Y) \sim P$. Using $(X_1, Y_1), \ldots, (X_n, Y_n)$ and X we want to predict Y. This is the same as regression except that Y is discrete.

We will focus on the simple case where $Y_i \in \{0, 1\}$. A *classifer* is a function h such that $h(x) \in \{0, 1\}$. The *risk* is

$$R(h) = P(Y \neq h(X)).$$

An estimate of the risk is the *training error* or *empirical risk*

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

1 Basic Theory

The best classifier is the *Bayes classifier* defined by:

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

where $m(x) = \mathbb{E}(Y|X = x)$. (Despite the name, this has nothing to do with Bayesian inference.)

Theorem 1 *For any h, $R(h) \geq R(h_*)$.*

Proof. For any h,

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

We will show that

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

for all x. Now

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

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

The Bayes classifier has a decision boundary defined by

$$D(h_*) = \{x : P(Y=1|X=x) = P(Y=0|X=x)\}.$$

We can also write

$$h_*(x) = \begin{cases} 1 & \text{if } \pi p_1(x) > (1-\pi)p_0(x) \\ 0 & \text{otherwise.} \end{cases}$$

where $p_0(x) = p(x|Y=0)$ and $p_1(x) = p(x|Y=1)$.

There are three main strategies for classification:

1. Empirical Risk Minimization: Choose a set of classifiers $\mathcal{H}$ and choose $\widehat{h}$ to minimize $\widehat{R}(h)$.

2. Regression or plug-in classification: Define

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{m}(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

   where $\widehat{m}(x)$ is an estimate of $m(x) = \mathbb{E}[Y|X=x] = P(Y=1|X=x)$.

3. Density Estimation: We estimate π, $p_0(x)$ and $p_1(x)$ using the training data, and then use the classifier:

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{\pi}\widehat{p}_1(x) > (1-\widehat{\pi})\widehat{p}_0(x) \\ 0 & \text{otherwise.} \end{cases}$$

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

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

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

Proof. We showed earlier that

$$R(h_g) - R(h_*) = \int \left[\mathbb{P}(Y \neq h_g(x)|X = x) - \mathbb{P}(Y \neq h_*(x)|X = x)\right] dP(x)$$

and that

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

Now

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

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

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

where the last step 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.

2 Linear Discriminant Analysis

Our first classifier to consider, will be based on density estimation. First, suppose that $p_0(x) = N(\mu_0, \Sigma)$ and $p_1(x) = N(\mu_1, \Sigma)$. In this simplified setting we can derive the form of the Bayes classifier. In particular, $h_*(x) = 1$ if:

$$\pi_1 \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2}\right) > (1-\pi_1)\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2}\right),$$

rearranging this we obtain that $h_*(x) = 1$ if,

$$\log(\pi_1/(1-\pi_1)) - \frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2} > -\frac{(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2}.$$

We note that the decision boundary of this classifier is:

$$\log(\pi_1/(1-\pi_1)) - \frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2} = -\frac{(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2},$$

which on re-arrangement gives:

$$\alpha_0 + \alpha_1^T x = 0$$

where

$$\begin{aligned}
\alpha_0 &= \log(\pi_1/(1-\pi_1)) - \frac{\mu_1^T\Sigma^{-1}\mu_1 + \mu_0^T\Sigma^{-1}\mu_0}{2} \\
\alpha_1 &= \Sigma^{-1}(\mu_1 - \mu_0)
\end{aligned}$$

which shows that the decision boundary of the classifier is linear. This is why the classifier is called *linear discriminant analysis*.

We can approximate the Bayes rule by estimating the various unknown quantities. Given a training data set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ we can estimate:

$$\begin{aligned}
\widehat{\pi}_1 &= \frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i = 1) \\
\widehat{\mu}_0 &= \frac{1}{\sum_{i=1}^n \mathbb{I}(Y_i = 0)}\sum_{i=1}^n X_i\mathbb{I}(Y_i = 0) \\
\widehat{\mu}_1 &= \frac{1}{\sum_{i=1}^n \mathbb{I}(Y_i = 1)}\sum_{i=1}^n X_i\mathbb{I}(Y_i = 1).
\end{aligned}$$

These are the maximum likelihood estimators for these parameters. The MLE for Σ is given by:

$$\begin{aligned}
\widehat{\Sigma}_0 &= \frac{1}{\sum_{i=1}^n \mathbb{I}(Y_i = 0)}\sum_{i=1}^n (X_i - \widehat{\mu}_0)(X_i - \widehat{\mu}_0)^T\mathbb{I}(Y_i = 0) \\
\widehat{\Sigma}_1 &= \frac{1}{\sum_{i=1}^n \mathbb{I}(Y_i = 1)}\sum_{i=1}^n (X_i - \widehat{\mu}_1)(X_i - \widehat{\mu}_1)^T\mathbb{I}(Y_i = 1) \\
\widehat{\Sigma} &= \frac{\sum_{i=1}^n \mathbb{I}(Y_i = 0)\widehat{\Sigma}_0 + \sum_{i=1}^n \mathbb{I}(Y_i = 1)\widehat{\Sigma}_1}{n}.
\end{aligned}$$

With these estimates in place we just use the rule $\widehat{h}(x) = 1$ if

$$\log(\widehat{\pi}_1/(1-\widehat{\pi}_1)) - \frac{\widehat{\mu}_1^T\widehat{\Sigma}^{-1}\widehat{\mu}_1 + \widehat{\mu}_0^T\widehat{\Sigma}^{-1}\widehat{\mu}_0}{2} + x^T\widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_0) > 0,$$

3 Logistic Regression

A popular direct regression based classifier is logistic regression where we assume that

$$m(x) = P(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}.$$

This is a logistic function of $\beta_0 + \beta^T x$ and has the property that it is always between $[0, 1]$.

Under the logistic hypothesis we can again derive the Bayes rule, $h_*(x)$ is 1 if:

$$\frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)} > \frac{1}{1 + \exp(\beta_0 + \beta^T x)},$$

which on rearrangement gives:

$$\beta_0 + \beta^T x > 0.$$

The decision boundary for the Bayes classifier is then simply:

$$\beta_0 + \beta^T x = 0,$$

which is again a linear decision boundary. So both LDA and logistic regression are linear classifiers.

To fit a logistic regression model, we maximize the conditional likelihood, i.e. we observe samples of the form $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we maximize,

$$\begin{aligned}\mathcal{L}(\beta_0, \beta) &= \prod_{i=1}^{n} P(Y_i = 1|X_i) \\ &= \prod_{i=1}^{n} \left(\frac{\exp(\beta_0 + \beta^T X_i)}{1 + \exp(\beta_0 + \beta^T X_i)}\right)^{Y_i} \left(\frac{\exp(\beta_0 + \beta^T X_i)}{1 + \exp(\beta_0 + \beta^T X_i)}\right)^{1-Y_i}.\end{aligned}$$

Unlike in linear regression, we cannot maximize the likelihood in closed form, so instead we use a method like gradient ascent. It turns out that the log-likelihood is a concave function so we can find the $(\widehat{\beta}_0, \widehat{\beta})$ that maximize the log-likelihood.

Connection Between Logistic Regression and LDA. We have seen that in both LDA and logistic regression our decision boundary is linear, i.e. we declare that $Y = 1$ if

$$\alpha_0 + \alpha^T X \geq 0,$$

for some scalar α_0 and vector $\alpha \in \mathbb{R}^d$. There is however, an important difference between LDA and logistic regression. In the two cases, we specify different likelihoods and fit the models differently.

In the LDA case, we used the joint likelihood:

$$\mathcal{L}(\beta_0, \beta) = \prod_{i=1}^{n} p(X_i, Y_i) = \prod_{i=1}^{n} p(Y_i)p(X_i|Y_i),$$

where we assumed that the first term is Bernoulli, while the second term is Gaussian. On the other hand for logistic regression, we use the conditional likelihood:

$$\mathcal{L}(\beta_0, \beta) = \prod_{i=1}^{n} p(Y_i|X_i)$$

which we assumed was a logistic function. In this case, we do not even model the distribution of X. In machine learning parlance, classifiers like LDA are called *generative classifiers* while classifiers like logistic regression are called *discriminative classifiers*.

4 Empirical Risk Classification

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 3 *Let h_* be the best classifier in $\mathcal{H}$, that is, $R(h)$ is minimized by h_*. Suppose that*

$$P(\sup_{h \in \mathcal{H}} |\widehat{R}(h) - R(h)| > \epsilon) < \delta. \tag{1}$$

Then, with probability at least $1 - \delta$, $R(\widehat{h}) \leq R(h_*) + 2\epsilon$.

Proof. Let A be the event that $|\widehat{R}(h) - R(h)| \leq \epsilon$ for all $h \in \mathcal{H}$. By assumption, this event has probability at least $1 - \delta$. And when A is true we have that

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

■

It follows that empirical risk minimization works well if we can show that (1) holds which is known as a *uniform bound*. So now we turn to the question: how do we show that (1) holds? This is the subject of the next section.

5 Uniform Bounds

Recall that, if $X_1, \ldots, X_n \sim$ Bernoulli(p) and $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i$ then, from Hoeffding's inequality,

$$\mathbb{P}(|\widehat{p}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Now we want to make a stronger statement like (1).

Generally, we can state out goal as follows. Let $\mathcal{A}$ be a class of sets. We want a bound of the form

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq c_1\kappa(\mathcal{A})e^{-c_2 n\epsilon^2}$$

where $P_n(A) = n^{-1}\sum_{i=1}^n I(X_i \in A)$. Bounds like these are called *uniform bounds* since they hold uniformly over a class of functions or over a class of sets.

Finite Classes. Let $\mathcal{A} = \{A_1, \ldots, A_N\}$. We will make use of the *union bound*. Recall that

$$\mathbb{P}\left(B_1 \bigcup \cdots \bigcup B_N\right) \leq \sum_{j=1}^N \mathbb{P}(B_j).$$

Let B_j be the event that $|P_n(A_j) - P(A_j)| > \epsilon$. From Hoeffding's inequality, $\mathbb{P}(B_j) \leq 2e^{-2n\epsilon^2}$. Then

$$\begin{aligned}\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) &= \mathbb{P}\left(B_1 \bigcup \cdots \bigcup B_N\right) \\ &\leq \sum_{j=1}^N \mathbb{P}(B_j) \leq \sum_{j=1}^N 2e^{-n\epsilon^2} = 2Ne^{-2n\epsilon^2}.\end{aligned}$$

Thus we have shown that

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 2\kappa e^{-n\epsilon^2}$$

where $\kappa = |\mathcal{A}|$.

To extend these ideas to infinite classes like $\mathcal{A} = \{(-\infty, t] : \ t \in \mathbb{R}\}$ we need to introduce a few more concepts.

Shattering. Let $\mathcal{A}$ be a class of sets. Some examples are:

1. $\mathcal{A} = \{(-\infty, t] : \ t \in \mathbb{R}\}$.
2. $\mathcal{A} = \{(a, b) : \ a \leq b\}$.
3. $\mathcal{A} = \{(a, b) \cup (c, d) : \ a \leq b \leq c \leq d\}$.

4. $\mathcal{A}$ = all discs in $\mathbb{R}^d$.
5. $\mathcal{A}$ = all rectangles in $\mathbb{R}^d$.
6. $\mathcal{A}$ = all half-spaces in $\mathbb{R}^d = \{x : \ \beta^T x \geq 0\}$.
7. $\mathcal{A}$ = all convex sets in $\mathbb{R}^d$.

Let $F = \{x_1, \ldots, x_n\}$ be a finite set. Let G be a subset of F. Say that $\mathcal{A}$ **picks out** G if

$$A \cap F = G$$

for some $A \in \mathcal{A}$. For example, let $\mathcal{A} = \{(a, b) : \ a \leq b\}$. Suppose that $F = \{1, 2, 7, 8, 9\}$ and $G = \{2, 7\}$. Then $\mathcal{A}$ picks out G since $A \cap F = G$ if we choose $A = (1.5, 7.5)$ for example.

Let $S(\mathcal{A}, F)$ be the number of these subsets picked out by $\mathcal{A}$. Of course $S(\mathcal{A}, F) \leq 2^n$.

Example 4 *Let* $\mathcal{A} = \{(a, b) : \ a \leq b\}$ *and* $F = \{1, 2, 3\}$. *Then* $\mathcal{A}$ *can pick out:*

$$\emptyset, \ \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}.$$

So $s(\mathcal{A}, F) = 7$. *Note that* $7 < 8 = 2^3$. *If* $F = \{1, 6\}$ *then* $\mathcal{A}$ *can pick out:*

$$\emptyset, \ \{1\}, \{6\}, \{1, 6\}.$$

In this case $s(\mathcal{A}, F) = 4 = 2^2$.

We say that F is **shattered** if $s(\mathcal{A}, F) = 2^n$ where n is the number of points in F.

Let $\mathcal{F}_n$ denote all finite sets with n elements.

Define the **shatter coefficient**

$$s_n(\mathcal{A}) = \sup_{F \in \mathcal{F}_n} s(\mathcal{A}, F).$$

Note that $s_n(\mathcal{A}) \leq 2^n$.

The following theorem is due to Vapnik and Chervonenis. The proof is beyond the scope of the course. (If you take 10-702/36-702 you will learn the proof.)

Theorem 5 *Let $\mathcal{A}$ be a class of sets. Then*

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}}|P_n(A)-P(A)|>\epsilon\right)\leq 8\ s_n(\mathcal{A})\ e^{-n\epsilon^2/32}. \tag{2}$$

This partly solves one of our problems. But, how big can $s_n(\mathcal{A})$ be? Sometimes $s_n(\mathcal{A})=2^n$ for all n. For example, let $\mathcal{A}$ be all polygons in the plane. Then $s_n(\mathcal{A})=2^n$ for all n. But, in many cases, we will see that $s_n(\mathcal{A})=2^n$ for all n up to some integer d and then $s_n(\mathcal{A})<2^n$ for all $n>d$.

Example 6 *Let $\mathcal{A}=\{(a,b):\ a,b\in\mathbb{R},a\leq b\}$. Then we have:*

n	2^n	s_n
1	*2*	*2*
2	*4*	*4*
3	*8*	*7*
4	*16*	*11*
$\vdots$	$\vdots$	$\vdots$

So $s_n=2^n$ for $n=1,2$. For $n>2$ we have $s_N<2^n$.

The **Vapnik-Chervonenkis (VC) dimension** is

$$d=d(\mathcal{A})=\quad\text{largest n such that } s_n(\mathcal{A})=2^n.$$

In other words, d is the size of the largest set that can be shattered.

Thus, $s_n(\mathcal{A})=2^n$ for all $n\leq d$ and $s_n(\mathcal{A})<2^n$ for all $n>d$. The VC dimensions of some common examples are summarized in Table 1. Now here is an interesting question: for $n>d$ how does $s_n(\mathcal{A})$ behave? It is less than 2^n but how much less?

Class $\mathcal{A}$	VC dimension $V_{\mathcal{A}}$
$\mathcal{A} = \{A_1, \ldots, A_N\}$	$\leq \log_2 N$
Intervals $[a, b]$ on the real line	2
Discs in $\mathbb{R}^2$	3
Closed balls in $\mathbb{R}^d$	$\leq d + 2$
Rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d + 1$
Convex polygons in $\mathcal{R}^2$	∞
Convex polygons with d vertices	$2d + 1$

Table 1: The VC dimension of some classes $\mathcal{A}$.

Theorem 7 (Sauer's Theorem) *Suppose that $\mathcal{A}$ has finite VC dimension d. Then, for all $n \geq d$,*

$$s(\mathcal{A}, n) \leq (n + 1)^d. \tag{3}$$

Sauer's Theorem is very surprising. It says there is a phase transition from exponential to polynomial. We conclude that:

Theorem 8 *Let $\mathcal{A}$ be a class of sets with VC dimension $d < \infty$. Then*

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 8\ (n + 1)^d\ e^{-n\epsilon^2/32}. \tag{4}$$

Theorem 9 *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 $\epsilon > 0$,

$$\mathbb{P}(R(\widehat{h}) > R(h_*) + 2\epsilon) \leq 8(n + 1)^d e^{-n\epsilon^2/32}$$

for some constants c_1 and c_2.

Proof. Recall that

$$\mathbb{P}(\sup_{h \in \mathcal{H}} |\widehat{R}(h) - R(h)| > \epsilon) \leq 8(n + 1)^d e^{-n\epsilon^2/32}.$$

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

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

■

This shows that empirical risk classification works if the VC dimension of $\mathcal{H}$ is finite.

6 Surrogate Loss Functions

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.* Logistic regression can also be seen as a surrogate loss function. Boosting is a method that uses the surrogate loss $e^{Yf(X)}$.

7 Nonparametric Classifiers

There are many approaches to nonparametric classification.

Kernel. Let $\widehat{m}$ be the kernel regression estimator. Then we simply set

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{m}(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Nearest Neighbors. Let $N(x)$ denote the k-nearest neighbors of x. Let $\widehat{m}(x) = (1/k)\sum_{N(x)} Y_i$ be the average value of the corresponding Y_i's. Then the k-nn classifier is defined by as above.

Naive Bayes. Another popular generative method is called Naive Bayes. We assume that

$$p(x|y) = \prod_{j=1}^{d} p_j(x(j)|y),$$

where we have written $x = (x(1), \ldots, x(j))$. Then we use one-dimensional non-parametric density estimation to estimate the densities $p_j(x(j)|y)$. We estimate the probability $P(Y = 1)$ as before and then classify a new point as belonging to class 1 if:

$$\widehat{\pi} \prod_{j=1}^{d} \widehat{p}_j(X(j)|Y = 1) \geq (1 - \widehat{\pi}) \prod_{j=1}^{d} \widehat{p}_j(X(j)|Y = 0).$$

8 Other Classifiers

There are of course many other classifiers including: trees, forests, boosting, deep neural nets, and so on. Currently, many of these methods do not have any theoretical support but they seem to work well in practice.