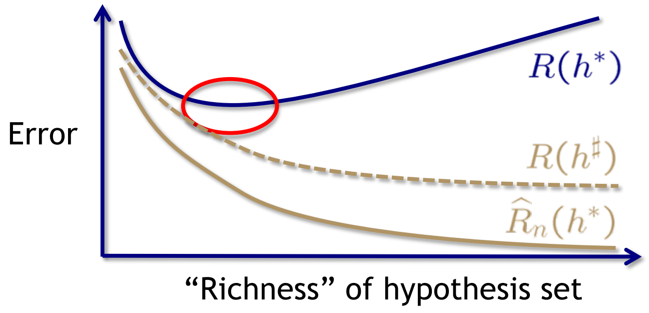
Fundamental tradeoff

Recall the definitions

$$h^\sharp = \arg\min_{h_j \in \mathcal{H}} R(h_j)$$

$$h^* = \arg\min_{h_j \in \mathcal{H}} \widehat{R}_n(h_j)$$



What is a good hypothesis?

Ideally, we would like to have a small number of hypotheses, so that $\widehat{R}_n(h^*) \approx R(h^*)$, while also being lucky (or smart) enough to have $R(h^*) \approx R(h^\sharp) \approx 0$

In general, this may not be possible, since there may not be ***any*** function h with $R(h) \approx 0$

Why not?

Noise: $Y = h(X) + N$

Suppose we knew the joint distribution of our data

- what is the optimal classification rule h^*?
- what are the fundamental limits on how small $R(h^*)$ can be?

Known distribution case

Consider (X, Y) where

- X is a random vector in $\mathbb{R}^d$
- $Y \in \{0, \ldots, K-1\}$ is a random variable (depending on X)

Let $h : \mathbb{R}^d \to \{0, \ldots, K-1\}$ be a ***classifier*** with ***probability of error/risk*** given by

$$R(h) := \mathbb{P}[h(X) \neq Y]$$

Our goal is to formulate a simple rule for minimizing $R(h)$ when the joint distribution of (X, Y) is known

Characterizations of the joint distribution

Let $f_{X,Y}(\mathbf{x}, k)$ denote the joint distribution of (X, Y)

- We can factor this as $f_{X,Y}(\mathbf{x}, k) = \pi_k f_{X|Y}(\mathbf{x}|k)$
  - $\pi_k = \mathbb{P}[Y = k]$ is the ***a priori probability*** of class k
  - $f_{X|Y}(\mathbf{x}|k)$ is the class conditional pdf (the pdf of $X|Y = k$)
- Alternatively (or via Bayes rule), we can factor this as $f_{X,Y}(\mathbf{x}, k) = \eta_k(\mathbf{x}) f_X(\mathbf{x})$
  - $\eta_k(\mathbf{x}) = \mathbb{P}[Y = k | X = \mathbf{x}]$ is the ***a posteriori probability*** of class k
  - $f_X(\mathbf{x})$ is simply the marginal pdf of X

The Bayes classifier

Theorem
The classifier $h^*(\mathbf{x}) := \arg\max_k \eta_k(\mathbf{x})$ satisfies

$$R^* = R(h^*) \le R(h)$$

for any possible classifier h

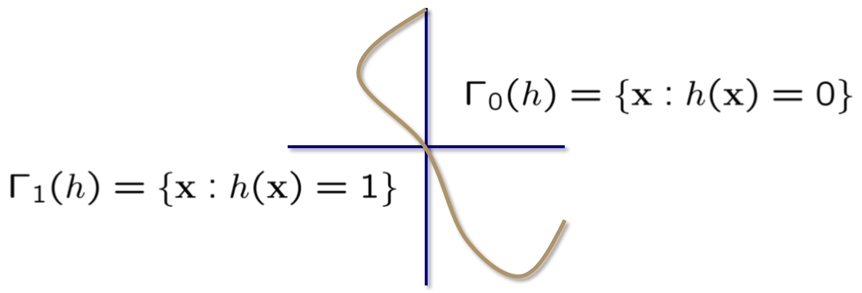
Terminology:
- h^* is called a ***Bayes classifier***
- R^* is called the ***Bayes risk***

Proof

For convenience, assume $X|Y = k$ is a continuous random variable with density $f_{X|Y}(\mathbf{x}|k)$

Consider an arbitrary classifier h. Denote the decision regions

$$\Gamma_k(h) := \{\mathbf{x} : h(\mathbf{x}) = k\}$$


$\Gamma_0(h) = \{\mathbf{x} : h(\mathbf{x}) = 0\}$
$\Gamma_1(h) = \{\mathbf{x} : h(\mathbf{x}) = 1\}$

Proof (Part 2)

We can write
$$\begin{aligned} 1 - R(h) &= \mathbb{P}[h(X) = Y] \\ &= \sum_{k=0}^{K-1} \pi_k \cdot \mathbb{P}[h(X) = k | Y = k] \\ &= \sum_{k=0}^{K-1} \pi_k \cdot \int_{\Gamma_k(h)} f_{X|Y}(\mathbf{x}|k)\, d\mathbf{x} \end{aligned}$$

We want to ***maximize*** this expression, we should design our classifier h such that

$$\mathbf{x} \in \Gamma_k(h) \quad \Longleftrightarrow \quad \pi_k f_{X|Y}(\mathbf{x}|k) \text{ is maximal}$$

Proof (Part 3)

Therefore, the optimal h has
$$\begin{aligned} h^*(\mathbf{x}) &= \arg\max_k \pi_k f_{X|Y}(\mathbf{x}|k) \\ &= \arg\max_k \frac{\pi_k f_{X|Y}(\mathbf{x}|k)}{\sum_{\ell=0}^{K-1} \pi_\ell f_{X|Y}(\mathbf{x}|\ell)} \\ &= \arg\max_k \mathbb{P}[Y = k | X = \mathbf{x}] \end{aligned}$$

Bayes rule!

Note that in addition to our rigorous derivation, this classifier also coincides with "common sense"

Variations

Different ways of expressing the Bayes classifier

- $h^*(\mathbf{x}) = \arg\max_k \eta_k(\mathbf{x})$
- $h^*(\mathbf{x}) = \arg\max_k \pi_k f_{X|Y}(\mathbf{x}|k)$
- When $K = 2$

$$\frac{f_{X|Y}(\mathbf{x}|1)}{f_{X|Y}(\mathbf{x}|0)} \underset{1}{\overset{0}{\lessgtr}} \frac{\pi_0}{\pi_1}$$

likelihood ratio test

- When $\pi_0 = \pi_1 = \cdots = \pi_{K-1}$

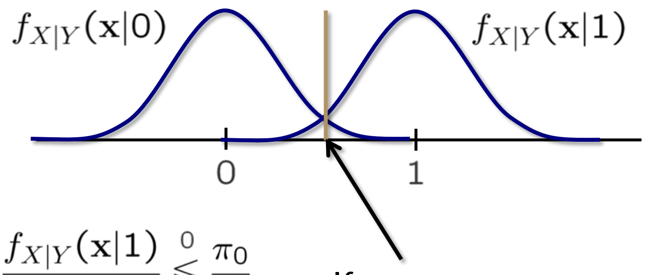
$$h^*(\mathbf{x}) = \arg\max_k f_{X|Y}(\mathbf{x}|k)$$

maximum likelihood classifier/detector

Example

Suppose that $K = 2$ and that

$$X|Y = 0 \sim \mathcal{N}(0,1) \qquad X|Y = 1 \sim \mathcal{N}(1,1)$$



$$\frac{f_{X|Y}(\mathbf{x}|1)}{f_{X|Y}(\mathbf{x}|0)} \underset{1}{\overset{0}{\lessgtr}} \frac{\pi_0}{\pi_1}$$

If $\pi_0 = \pi_1$

Example

How do we calculate the ***Bayes risk***?

$$\begin{aligned} R(h^*) &= \mathbb{P}\left[h^*(X) \neq Y\right] \\ &= \mathbb{P}\left[\text{declare } 0|Y=1\right] \cdot \pi_1 \\ &\quad + \mathbb{P}\left[\text{declare } 1|Y=0\right] \cdot \pi_0 \end{aligned}$$

In the case where $\pi_0 = \pi_1 = \frac{1}{2}$, our test reduced to declaring 1 iff $x \geq \frac{1}{2}$, thus

$$\begin{aligned} R(h^*) &= \tfrac{1}{2}\mathbb{P}\left[X < \tfrac{1}{2}|Y=1\right] + \tfrac{1}{2}\mathbb{P}\left[X > \tfrac{1}{2}|Y=0\right] \\ &= \frac{1}{2}\int_{-\infty}^{\frac{1}{2}} f_{X|Y}(x|1)\,dx + \frac{1}{2}\int_{\frac{1}{2}}^{\infty} f_{X|Y}(x|0)\,dx \\ &= \Phi\left(-\tfrac{1}{2}\right) \end{aligned}$$

Alternative cost/loss functions

So far we have focused on minimizing the risk $\mathbb{P}[h(X) \neq Y]$

There are many situations where this is not appropriate

- cost-sensitive classification
  - type I/type II errors or misses/false alarms may have very different costs, in which case a loss function of the form

$$C_0\mathbb{P}[h(X) \neq Y|Y=0] + C_1\mathbb{P}[h(X) \neq Y|Y=1]$$

  - alternatively, it may be better to focus on them directly a la Neyman-Pearson classification
- unbalanced datasets
  - when one class dominates the other, the probability of error will place less emphasis on the smaller class
  - the class proportions in our dataset may not be representative of the "wild"

What about learning?

We have just seen that when we know the true distribution underlying our dataset, solving the classification problem is straightforward

In practical learning problems, all we have is the data…

One natural approach is to use the data to estimate the distribution, and then just plug this into the formula for the Bayes classifier

Plugin methods

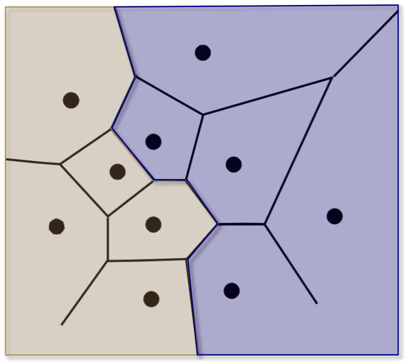
Before we get to these, we will first talk about what is quite possibly the absolute simplest learning algorithm there is…

Nearest neighbor classifier

The ***nearest neighbor classifier*** is easiest to state in words:

Assign $\mathbf{x}$ the same label as the closest training point $\mathbf{x}_i$ to $\mathbf{x}$

The nearest neighbor rule defines a ***Vornoi partition*** of the input space



Risk of the nearest neighbor classifier

To simplify our discussion, we will restrict our attention to the binary case where $Y \in \{0,1\}$

Consider the Bayes risk conditioned on $X = \mathbf{x}$:

$$R^*(\mathbf{x}) := \mathbb{P}[Y \neq h^*(\mathbf{x}) | X = \mathbf{x}]$$

Note that if $h^*(\mathbf{x}) = 0$, then we must have $R^*(\mathbf{x}) = \eta_1(\mathbf{x})$

Similarly, if $h^*(\mathbf{x}) = 1$, then $R^*(\mathbf{x}) = \eta_0(\mathbf{x})$

Since $h^*(\mathbf{x})$ selects the label that maximizes $\eta_k(\mathbf{x})$, we thus have that

$$R^*(\mathbf{x}) = \min\left(\eta_0(\mathbf{x}), \eta_1(\mathbf{x})\right)$$

Risk of the nearest neighbor classifier

Now consider the risk of the nearest neighbor classifier conditioned on $X = \mathbf{x}$

$$R^{\mathrm{NN}}(\mathbf{x}) := \mathbb{P}[Y \neq h^{\mathrm{NN}}(\mathbf{x}) | X = \mathbf{x}]$$

Note that for a fixed $\mathbf{x}$, we are treating Y as random

Here we will further treat $h^{\mathrm{NN}}(\mathbf{x})$ as being random since ***it depends on the training dataset***

Thus we have that

$$R^{\mathrm{NN}}(\mathbf{x}) = \mathbb{P}\left[Y = 0 | X = \mathbf{x}\right] \mathbb{P}\left[h^{\mathrm{NN}}(\mathbf{x}) = 1 | X = \mathbf{x}\right]$$
$$+\mathbb{P}\left[Y = 1 | X = \mathbf{x}\right] \mathbb{P}\left[h^{\mathrm{NN}}(\mathbf{x}) = 0 | X = \mathbf{x}\right]$$

Risk of the nearest neighbor classifier

$$R^{\text{NN}}(\mathbf{x}) = \mathbb{P}[Y=0|X=\mathbf{x}]\,\mathbb{P}\left[h^{\text{NN}}(\mathbf{x})=1|X=\mathbf{x}\right] + \mathbb{P}[Y=1|X=\mathbf{x}]\,\mathbb{P}\left[h^{\text{NN}}(\mathbf{x})=0|X=\mathbf{x}\right]$$

Note that if $\mathbf{x}_{\text{NN}}$ is the nearest neighbor to $\mathbf{x}$, then

$$\mathbb{P}\left[h^{\text{NN}}(\mathbf{x})=k|X=\mathbf{x}\right] = \mathbb{P}[Y=k|X=\mathbf{x}_{\text{NN}}] = \eta_k(\mathbf{x}_{\text{NN}})$$

Thus, we can write

$$R^{\text{NN}}(\mathbf{x}) = \eta_0(\mathbf{x})\eta_1(\mathbf{x}_{\text{NN}}) + \eta_1(\mathbf{x})\eta_0(\mathbf{x}_{\text{NN}})$$

Intuition from asymptotics

In the limit as $n \to \infty$, we can assume that $\mathbf{x}_{\text{NN}} \to \mathbf{x}$

Thus, as $n \to \infty$ we have

$$R^{\text{NN}}(\mathbf{x}) \to \eta_0(\mathbf{x})\eta_1(\mathbf{x}) + \eta_1(\mathbf{x})\eta_0(\mathbf{x}) = 2\eta_0(\mathbf{x})\eta_1(\mathbf{x})$$

It is easy to see that

$$2\eta_0(\mathbf{x})\eta_1(\mathbf{x}) \le 2\min\left(\eta_0(\mathbf{x}),\eta_1(\mathbf{x})\right) = 2R^*(\mathbf{x})$$

Asymptotically, the risk of the nearest neighbor classifier is ***at most twice the Bayes risk***

k-nearest neighbors

We can drive the factor of 2 in this result down to 1 by generalizing the nearest neighbor rule to the ***k-nearest neighbor*** rule as follows:

Assign a label to $\mathbf{x}$ by taking a majority vote over the k training points $\mathbf{x}_i$ closest to $\mathbf{x}$

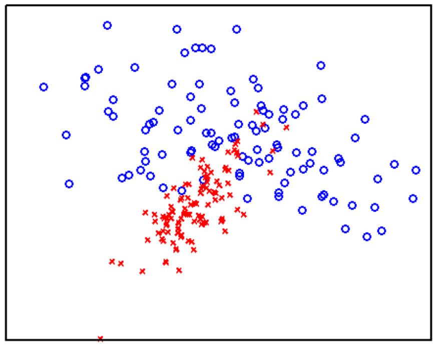
How do we define this more mathematically?

$I_k(\mathbf{x}) :=$ indices of the k training points closest to $\mathbf{x}$

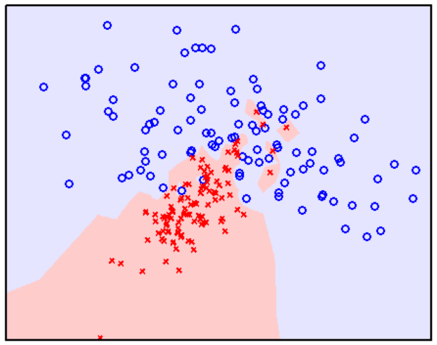
If $y_i = \pm 1$, then we can write the k-nearest neighbor classifier as

$$\widehat{h}_k(\mathbf{x}) := \text{sign}\left(\sum_{i \in I_k(\mathbf{x})} y_i\right)$$

Example

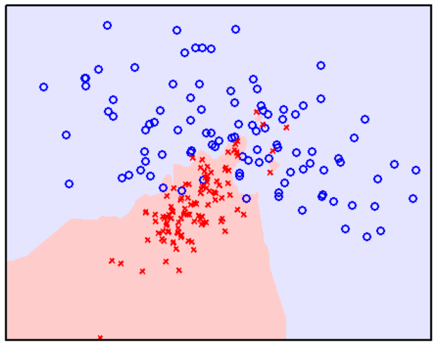


Example



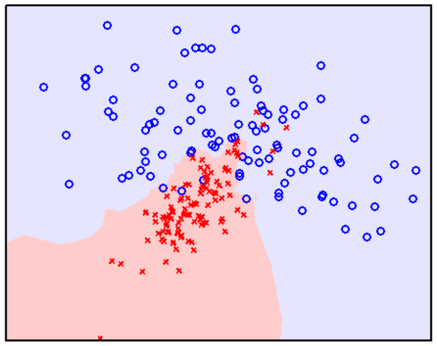
$k = 1$

Example



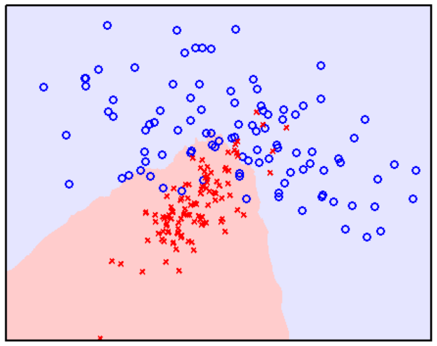
$k = 3$

Example



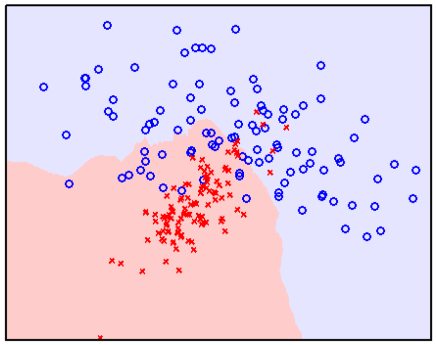
$k = 5$

Example



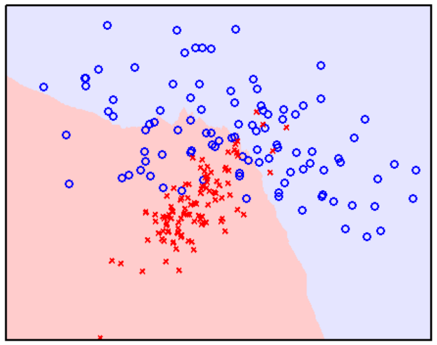
$k = 25$

Example



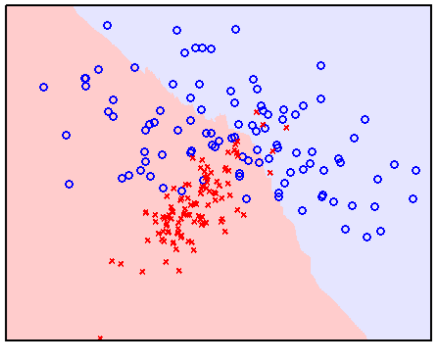
$k = 51$

Example



$k = 75$

Example



$k = 101$

Choosing the size of the neighborhood

Setting the parameter k is a problem of ***model selection***

Setting k by trying to minimize the training error is a particularly bad idea

$$\widehat{R}_n(\widehat{h}_k) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\widehat{h}_k(\mathbf{x}_i) \neq y_i\}}$$

What is $\widehat{R}_n(\widehat{h}_1)$?

No matter what, we always have $\widehat{R}_n(\widehat{h}_1) = 0$

Not much practical guidance from the theory, so we typically must rely on estimates based on holdout sets or more sophisticated model selection techniques

Consistency

We say that a classification algorithm is ***consistent*** if when the size of the training set $n \to \infty$, we have that

$$\mathbb{E}\left[R(\widehat{h}_n)\right] \to R^*$$

where $\widehat{h}_n$ is a classifier learned from a training set of size n and R^* is the Bayes risk

Theorem

Let $\widehat{h}_{k,n}$ denote the k-nearest neighbor rule with a training set of size n. If $n \to \infty$, $k \to \infty$, and $k/n \to 0$, then $\widehat{h}_{k,n}$ is consistent, i.e.,

$$\mathbb{E}\left[R(\widehat{h}_{k,n})\right] \to R^*$$

Summary

Given enough data, the k-nearest neighbor classifier will do just as well as pretty much any other method

Catch

- The amount of required data can be huge, especially if our feature space is high-dimensional
- The parameter k can matter a lot, so model selection will can be very important
- Finding the nearest neighbors out of a set of millions of examples is still pretty hard
  - can be sped up using k-d trees, but can still be relatively expensive to apply
  - in contrast, many of the other algorithms we will study have an expensive "training" phase, but application is cheap