Classification Problem

- Given \( \{(x_n, y_n)\}_{n=1}^{N} \) \( x_n \in \mathbb{R}^q \) \( y_n \in \{+, -\} \)

- Predict class label of a given query \( x_0 \)

Classification Problem

- Unknown probability distribution \( P(x, y) \)

- We need to estimate:

\[
P(+) \mid x_0 \equiv f_+ (x_0)
\]

\[
P(-) \mid x_0 \equiv f_- (x_0)
\]
The Bayesian Classifier

- Loss function: $\lambda(j \mid k)$
- Expected loss (conditional risk) associated with class $j$:
  \[ R(j \mid x) = \sum_{k=1}^{J} \lambda(j \mid k) P(k \mid x) \]
- Bayes rule:
  \[ j^* = \arg \min_{1 \leq j \leq J} R(j \mid x) \]
- Zero-one loss function:
  \[ \lambda(j \mid k) = \begin{cases} 
  0 & \text{if } j = k \\
  1 & \text{if } j \neq k 
\end{cases} \]
  \[ j^* = \arg \max_{1 \leq j \leq J} P(j \mid x) \]

Bayes rule

The Bayesian Classifier

- Bayes rule achieves the minimum error rate
- How to estimate the posterior probabilities:
  \[ \{ P(j \mid x) \}_{j=1}^{J} \]
  \[ \hat{j}(x) = \arg \max_{1 \leq j \leq J} \hat{P}(j \mid x) \]
Density estimation

- Use Bayes theorem to estimate the posterior probability values:

\[
P(j \mid x) = \frac{p(x \mid j)P(j)}{\sum_{k=1}^{l} p(x \mid k)P(k)}
\]

- \(p(x \mid j)\) is the probability density function of \(x\) given class \(j\)

- \(P(j)\) is the prior probability of class \(j\)

Naïve Bayes Classifier

- Makes the assumption of independence of features given the class:

\[
p(x \mid j) = p(x_1, x_2, \ldots, x_q \mid j) = \prod_{i=1}^{q} p(x_i \mid j)
\]

- The task of estimating a \(q\)-dimensional density function is reduced to the estimation of \(q\) one-dimensional density functions. Thus, the complexity of the task is drastically reduced.

- The use of Bayes theorem becomes much simpler.

- Proven to be effective in practice.
Nearest-Neighbor Methods

• Predict the class label of $x_0$ as the most frequent one occurring in the $K$ neighbors.
Nearest-Neighbor Methods

- Predict the class label of \( x_0 \) as the most frequent one occurring in the \( K \) neighbors.

**Basic assumption:**

\[
\begin{align*}
  f_+(x + \delta x) &\approx f_+(x) \\
  f_-(x + \delta x) &\approx f_-(x)
\end{align*}
\]

for small \( \|\delta x\| \)

Example: Letter Recognition

First statistical moment: 

Edge count:
Asymptotic Properties of K-NN Methods

\[ \lim_{N \to \infty} \hat{f}_j(x) = f_j(x) \]

if \( \lim_{N \to \infty} K = \infty \) and \( \lim_{N \to \infty} K/N = 0 \)

- The first condition reduces the variance by making the estimation independent of the accidental characteristics of the \( K \) nearest neighbors.

- The second condition reduces the bias by assuring that the \( K \) nearest neighbors are arbitrarily close to the query point.

Asymptotic Properties of K-NN Methods

\[ \lim_{N \to \infty} E_1 \leq 2E_\infty \]

\( E_1 \) = classification error rate of the 1-NN rule
\( E_\infty \) = classification error rate of the Bayes rule

In the asymptotic limit no decision rule is more than twice as accurate as the 1-NN rule
Finite-sample settings

- How well the 1-NN rule works in finite-sample settings?

- If the number of training data $N$ is large and the number of input features $q$ is small, then the asymptotic results may still be valid.

- However, for a moderate to large number of input variables, the sample required for their validity is beyond feasibility.

Curse-of-Dimensionality

- This phenomenon is known as the **curse-of-dimensionality**

- It refers to the fact that in high dimensional spaces data become extremely sparse and are far apart from each other

- It affects any estimation problem with high dimensionality