Applied Statistical Multivariate Analysis

Classification Part II

Non-Parametric Classifiers- K Nearest Neighbors

Instructor: Yaser Zerehsaz

موسسه آموزش عالی آزاد توسعه

برگزار کننده دوره‌های تخصصی علم داده

Problem Definition

- In theory, we would like to estimate the conditional probability of y given $\mathbf{x}$

$$P_{y|\mathbf{x}} = \frac{P_{\mathbf{x}|y} P_y}{P_{\mathbf{x}}}$$

- In Bayesian methods, we assumed that $P_{\mathbf{x}|y}$ is a multivariate normal distribution
- However, this is not always the case in practice
- How can we estimate $P_{y|\mathbf{x}}$ without making any assumptions about $P_{\mathbf{x}|y}$

Density estimation

- Given $\mathbf{z}$ as a random vector, we would like to estimate the density $P(\mathbf{z})$.
- Given a region $\mathcal{R}$, the probability that $\mathbf{z}$ falls in this region is

$$P = \int_{\mathcal{R}} P(\mathbf{z}) d\mathbf{z}$$

- Now, suppose that we have N observations K of which falling in region $\mathcal{R}$
- K is a binomial random variable with a mass function given by

$$\binom{N}{K} P^K (1-P)^{N-K}$$

So, $E\left(\frac{K}{N}\right) = P$ is the mean fraction of observations falling in $\mathcal{R}$

and $var\left(\frac{K}{N}\right) = \frac{P(1-P)}{N}$ is the variability of this fraction

Density estimation

- What happens if $N \to \infty$?
- In this case, $var\left(\frac{K}{N}\right) = \frac{P(1-P)}{N} \to 0$ and $K \cong NP$
- Also, if we assume that $\mathcal{R}$ is small enough that $P(\mathbf{z})$ is constant over the region, we have

$$P = \int_{\mathcal{R}} P(\mathbf{z})d\mathbf{z} \cong P(\mathbf{z}) \int_{\mathcal{R}} d\mathbf{z} \cong P(\mathbf{z})V$$

  Where V is the volume of $\mathcal{R}$

- Hence, we can write

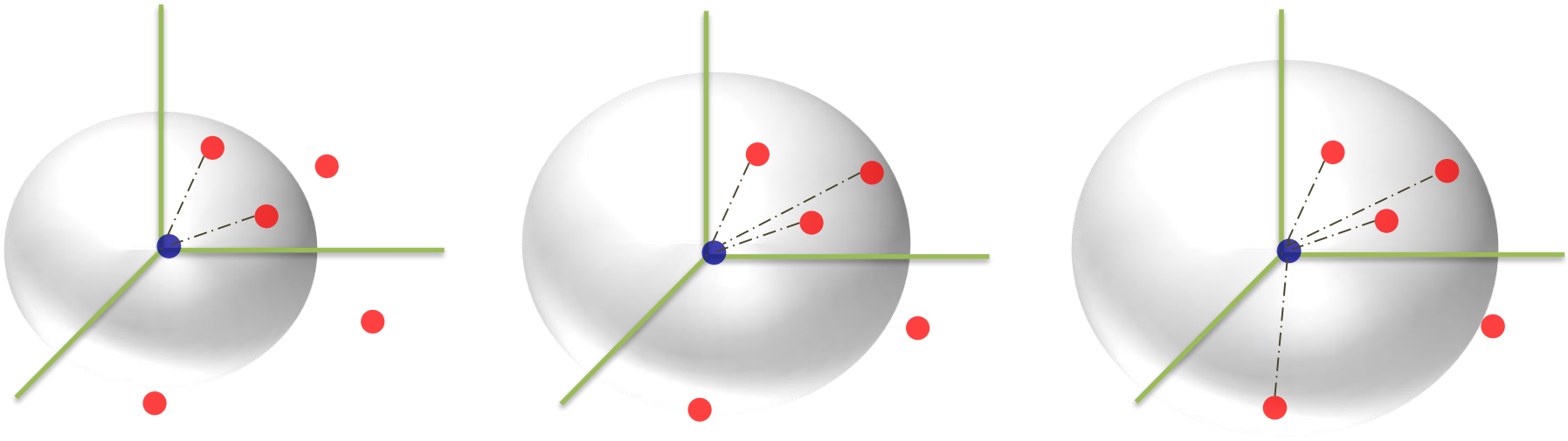
$$K \cong NP \cong NP(\mathbf{z})V$$

$$P(\mathbf{z}) \cong \frac{K}{NV}$$

- So, we can fix K, and determine the value of V from the data

K Nearest Neighbors Density Estimation

- In KNN, we consider a small sphere centered on $\mathbf{z}$
- We increase the radius of the sphere until it covers K points
- We let V be the volume of this sphere
- If the density is high around point $\mathbf{z}$, V will be small



- Increasing V to cover K points means finding K nearest neighbors

Back to the Classification Problem

- We would like to estimate $P_{y|\mathbf{x}}$, so we need to estimate $P_{\mathbf{x}|y}$
- Suppose that we have N observations with M classes
- Each class m has N_m observations and $\sum_{m=1}^{M} N_m = N$
- Now, we have a new observation $\mathbf{x}$ and we would like to classify it
- We find the K nearest neighbors of $\mathbf{x}$ irrespective of their classes

Back to the Classification Problem

- The conditional distribution of $\mathbf{x}$ given class $y = m$ is given as

$$P_{\mathbf{x}|y=m} = \frac{K_m}{N_m V}$$

- The marginal density of $\mathbf{x}$ can be defined as

$$P_{\mathbf{x}} = \frac{K}{NV}$$

- The posterior probability of $y = j$ given $\mathbf{x}$ is computed as

$$P_{y=m|\mathbf{x}} = \frac{P_{\mathbf{x}|m} P_m}{P_{\mathbf{x}}} = \frac{\frac{K_m}{N_m V} \times \frac{N_m}{N}}{\frac{K}{NV}} = \frac{K_m}{K}$$

- Now, we can apply the Bayes rule again as

$$f^* = \underset{m=1,2,\ldots,M}{\operatorname{argmax}} P_{y=m|\mathbf{x}} = \underset{m=1,2,\ldots,M}{\operatorname{argmax}} \frac{K_m}{K}$$

Types of Distances

- The distance between two points $\mathbf{x}$ and $\tilde{\mathbf{x}}$ can be measured in different ways
- Euclidean distance:

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_2 = \sqrt{\sum_{j=1}^{p} \left(x_j - \tilde{x}_j\right)^2}$$

- L_1 distance:

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_1 = \sum_{j=1}^{p} \left|x_j - \tilde{x}_j\right|$$

Selecting K

When K is small

- Our classifier is “more blind” to the overall distribution.
- Overly flexible fit, which will have low bias but high variance. Graphically, our decision boundary will be more jagged

When K is large

- More votes are averaged in each prediction and hence is more resilient to outliers.
- Larger values of K will have smoother decision boundaries which means lower variance but increased bias.

What happens if $K = N$?

Pros and Cons

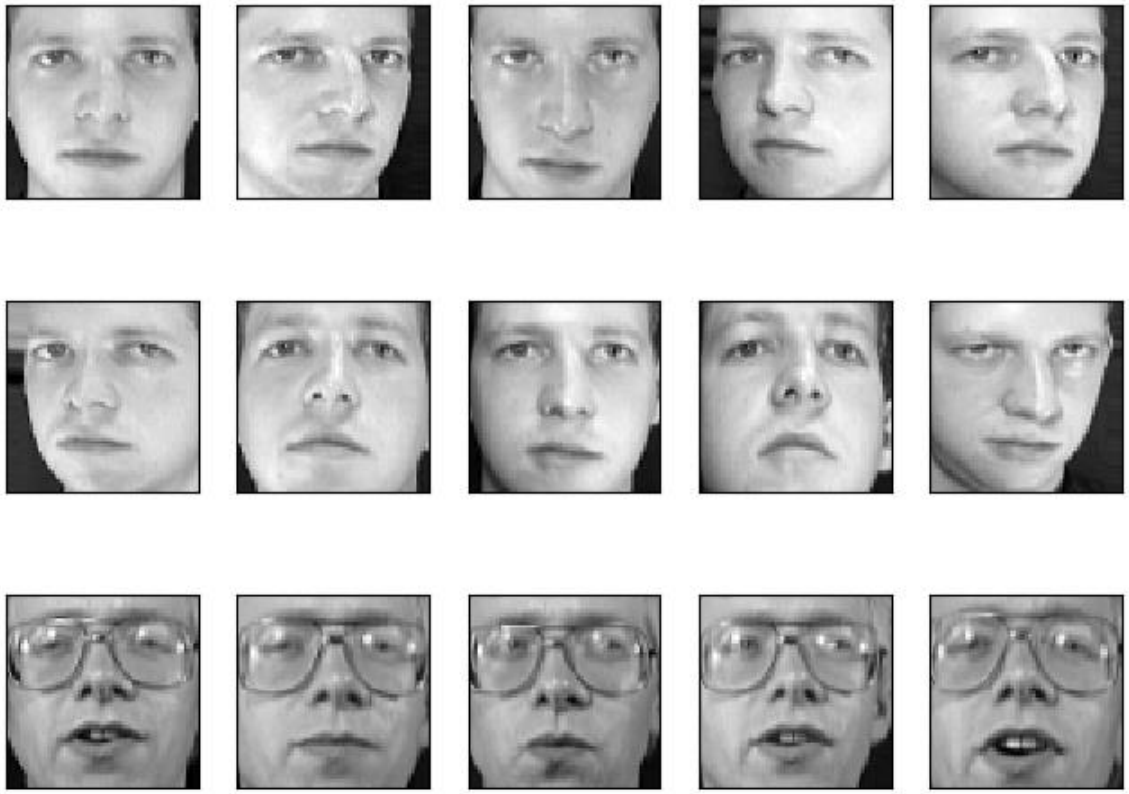
Cons

- The KNN algorithm is computationally expensive in the testing phase
- KNN can suffer from skewed class distributions. For example, if a certain class is very frequent in the training set, it will tend to dominate the majority voting of the new example

Pros

- KNN algorithm is simple to understand and easy to implement with zero to little training time

Face Recognition



Bias-Variance Tradeoff

- Lets assume that we have a training dataset $\mathcal{T}$ and $Y = f(x) + \varepsilon$
- We can use inputs X and labels Y to estimate a function (classifier) $f(x)$ with $\hat{f}(x)$
- For a squared error loss function, the expected test error for a new observation x_0

$$Err = E(Y - \hat{f}(x_0))^2 = E(Y - \hat{f}(x_0) + f(x) - f(x))^2 = \sigma_\varepsilon^2 + Bias^2\left(\hat{f}(x_0)\right) + var\left(\hat{f}(x_0)\right)$$

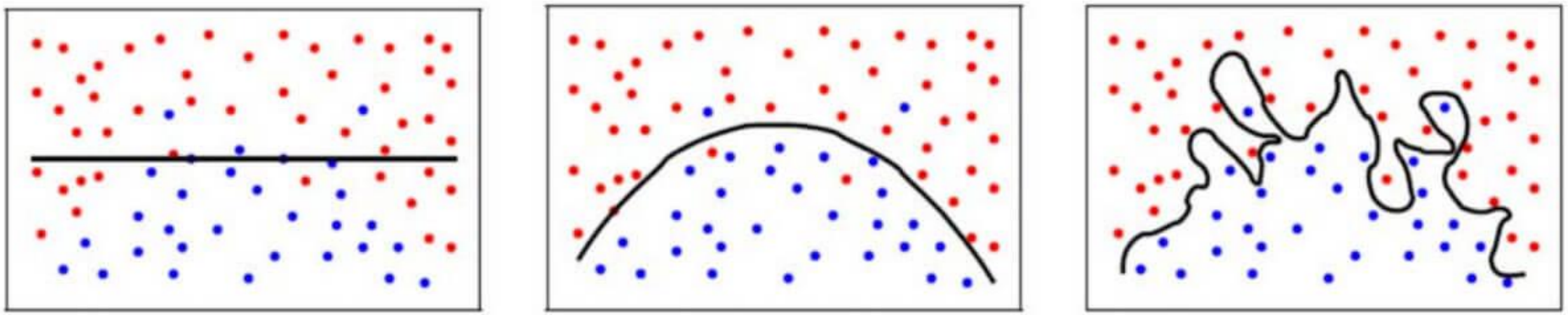
$$Bias = f(x) - E\left(\hat{f}(x_0)\right)$$

$$var\left(\hat{f}(x_0)\right) = E\left(\left[\hat{f}(x_0) - E\left(\hat{f}(x_0)\right)\right]^2\right)$$

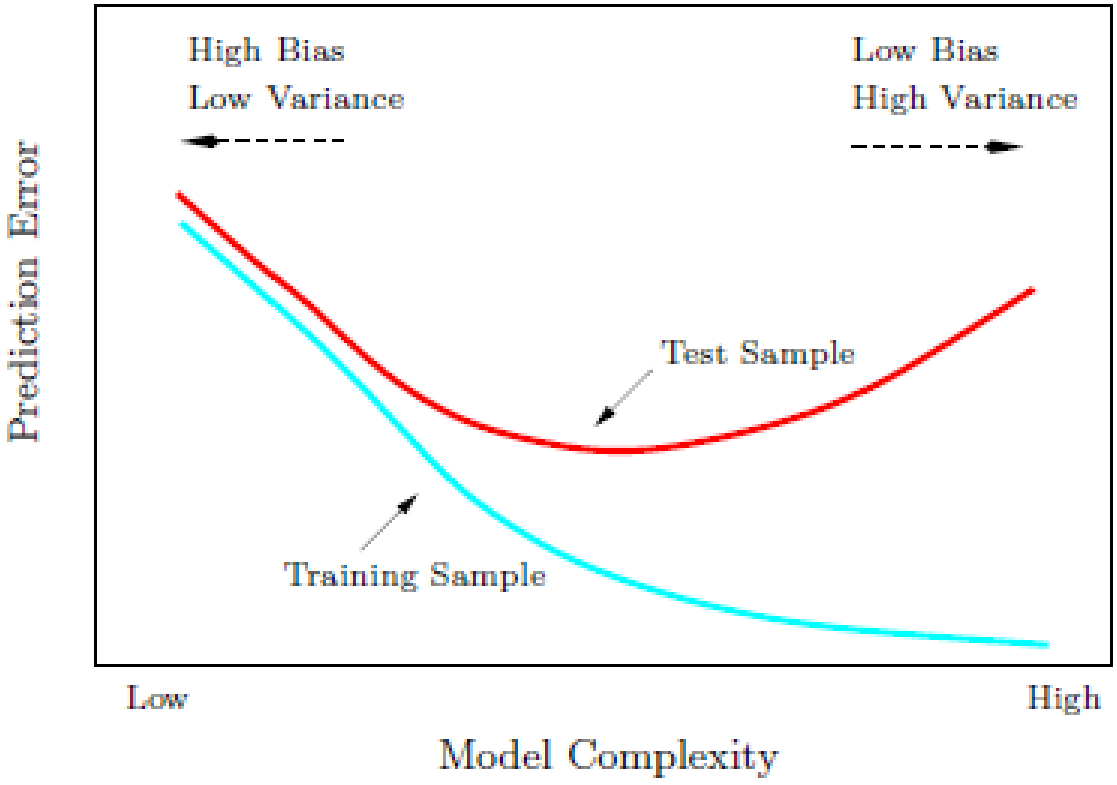
Bias-Variance Tradeoff

Generalization Problem in Classification

Underfitting ⟷ Overfitting



Bias-Variance Tradeoff



Hastie, Tibshirani and Friedman (2008)

Cross Validation

- Divide your dataset $\mathcal{D}$ into two groups of training and testing sets
- Divide your training dataset $\mathcal{T}$ into K folds (groups) with $K-1$ folds for training and one fold (kth fold) for validation

Train	Validation	Train	Train	Train	Train

- Train your model using $K-1$ training folds and compute the training error for these folds
- Compute the validation error using the kth fold
- Repeat this for $k = 1,2,\dots,K$

The cross validation error is the average of all K validation errors

The training error is the average of all K training errors

The testing error is computed based on the testing set

Select the tuning parameters minimizing cross validation error