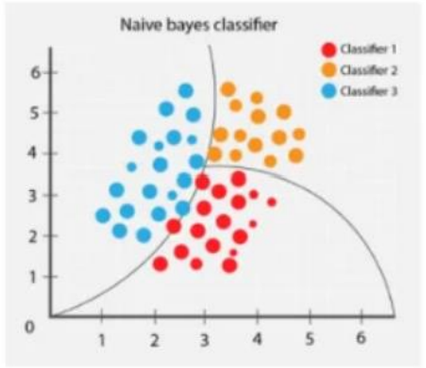
دسته‌بندی

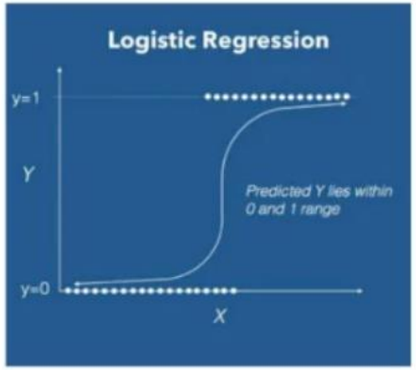
Classification

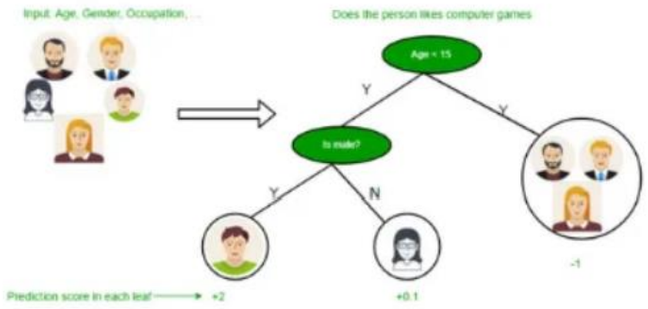
یک دسته‌بندی کننده، یک مدل یا تابعی است که به یک نقطه‌ی داده شده‌ی $\boldsymbol{x}$ دسته‌ای را تخصیص می‌دهد.

$$\hat{y} = M(\boldsymbol{x}), \qquad \hat{y} \in \{c_1, c_2, \dots, c_k\}, \qquad c_i \; is \; a \; class \; label$$

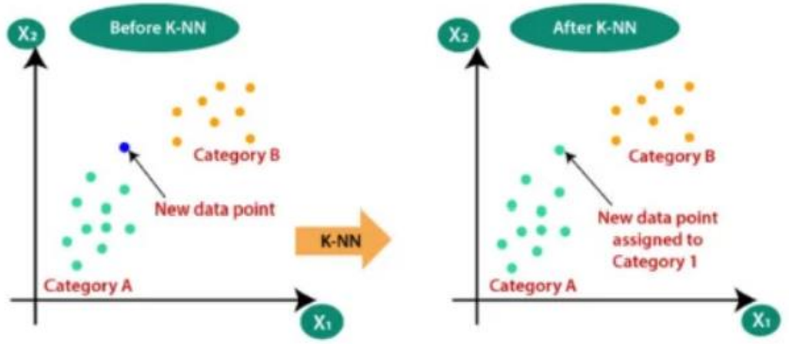
از یک مجموعه داده‌ی یادگیری (Training Set) استفاده می‌کنیم تا ماشین مدل $M(\boldsymbol{x})$ را بیاموزد.



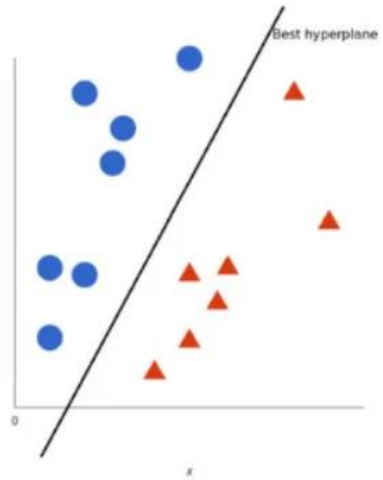




Decision Tree



K-Nearest Neighbors



Support Vector Machines

دسته‌بندی احتمالاتی

Probabilistic Classification

دسته‌بندی‌گر بیز(Bayes Classifier): مشخصه‌های عددی (Numerical Attributes)

$$\hat{y} = \arg\max_{c_i}\{P(c_i|\mathbf{x})\} = \arg\max_{c_i}\left\{\frac{P(\mathbf{x}|c_i)P(c_i)}{P(\mathbf{x})}\right\} \qquad P(\mathbf{x}) = \sum_{j=1}^{k} P(\mathbf{x}|c_j)\cdot P(c_j)$$

$$= \arg\max_{c_i}\left\{P(\mathbf{x}|c_i)P(c_i)\right\}$$

$$\hat{P}(c_i) = \frac{n_i}{n}$$

$$f_i(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(\sqrt{2\pi})^d\sqrt{|\boldsymbol{\Sigma}_i|}}\exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2}\right\}$$

$$P(c_i|\mathbf{x}) = \frac{2\epsilon\cdot f_i(\mathbf{x})P(c_i)}{\sum_{j=1}^{k} 2\epsilon\cdot f_j(\mathbf{x})P(c_j)} = \frac{f_i(\mathbf{x})P(c_i)}{\sum_{j=1}^{k} f_j(\mathbf{x})P(c_j)}$$

$$\hat{f}_i(\mathbf{x}) = f(\mathbf{x}|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)$$

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i}\sum_{\mathbf{x}_j\in\mathbf{D}_i}\mathbf{x}_j$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i}\bar{\mathbf{D}}_i^T\,\bar{\mathbf{D}}_i \qquad \bar{\mathbf{D}}_i = \mathbf{D}_i - \mathbf{1}\cdot\hat{\boldsymbol{\mu}}_i^T$$

دسته‌بندی‌گر بیز(Bayes Classifier): مشخصه‌های عددی (Numerical Attributes)

Algorithm 18.1: Bayes Classifier

BAYESCLASSIFIER (D):

1 **for** $i = 1, \ldots, k$ **do**
2 $\quad \mathbf{D}_i \leftarrow \{\mathbf{x}_j^T \mid y_j = c_i, j = 1, \ldots, n\}$ // class-specific subsets
3 $\quad n_i \leftarrow |\mathbf{D}_i|$ // cardinality
4 $\quad \hat{P}(c_i) \leftarrow n_i/n$ // prior probability
5 $\quad \hat{\boldsymbol{\mu}}_i \leftarrow \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{D}_i} \mathbf{x}_j$ // mean
6 $\quad \overline{\mathbf{D}}_i \leftarrow \mathbf{D}_i - \mathbf{1}_{n_i} \hat{\boldsymbol{\mu}}_i^T$ // centered data
7 $\quad \widehat{\boldsymbol{\Sigma}}_i \leftarrow \frac{1}{n_i} \overline{\mathbf{D}}_i^T \overline{\mathbf{D}}_i$ // covariance matrix
8 **return** $\hat{P}(c_i), \hat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i$ *for all* $i = 1, \ldots, k$

TESTING (x and $\hat{P}(c_i)$, $\hat{\boldsymbol{\mu}}_i$, $\widehat{\boldsymbol{\Sigma}}_i$, for all $i \in [1, k]$):

9 $\hat{y} \leftarrow \arg\max_{c_i} \{ f(\mathbf{x} | \hat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) \cdot P(c_i) \}$
10 **return** $\hat{y}$

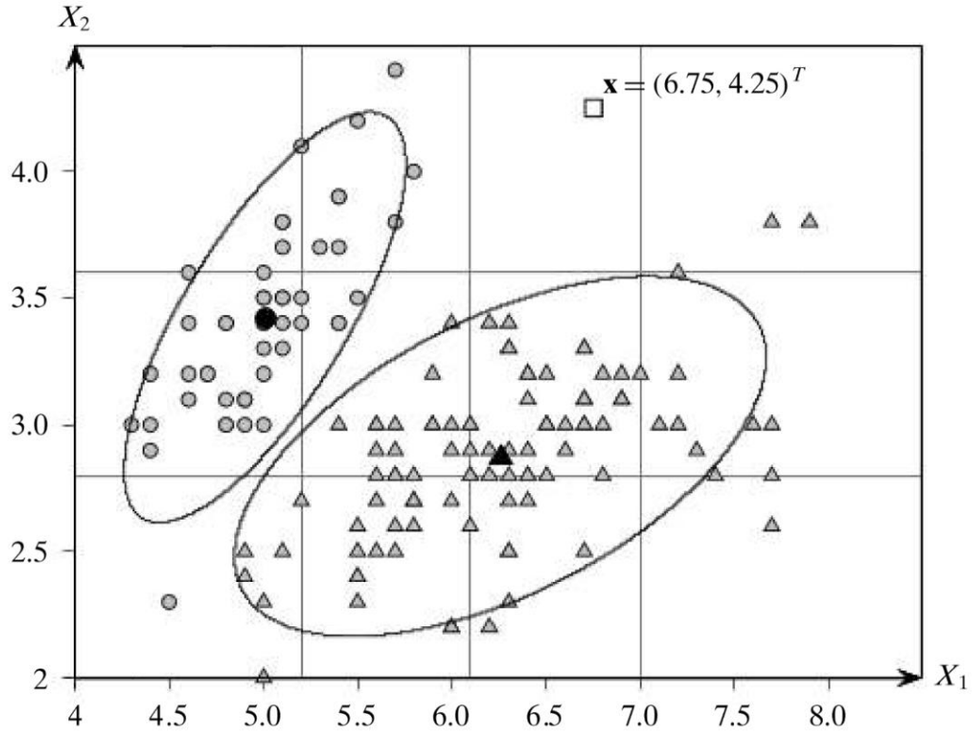


Figure 18.1. Iris data: X_1:sepallength versus X_2:sepal width. The class means are show in black; the density contours are also shown. The square represents a test point labeled **x**.

دسته‌بندی‌گر بیز(Bayes Classifier): مشخصه‌های دسته‌ای (Categorical Attributes)

متغیر X_j تعداد m_j مقدار متفاوت می‌گیرد، پس می‌توان با یک متغیر برنولی m_j–بعدی مدل نمود که آن را با $\boldsymbol{X}_j$ نشان می‌دهیم و به ازای هر مقدار، یکی از بردارهای $\boldsymbol{e}_{j1}, \boldsymbol{e}_{j2}, \ldots, \boldsymbol{e}_{jm_j} \in \mathbb{R}^{m_j}$ را می‌گیرد.

$$dom(X_j) = \{a_{j1}, a_{j2}, \ldots, a_{jm_j}\}$$

می‌توان همه‌ی d متغیر X_j را که هر یک m_j–بعد دارد در یک بردار binary به ابعاد $d' = \sum_{j=1}^{d} m_j$ آورد.

$$\boldsymbol{v} = \begin{pmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_d \end{pmatrix} = \begin{pmatrix} \boldsymbol{e}_{1r_1} \\ \vdots \\ \boldsymbol{e}_{dr_d} \end{pmatrix}, \qquad \boldsymbol{v} \in \mathbb{R}^{d'}$$

درست‌نمایی و احتمال پیشین از روابط زیر محاسبه می‌شوند.
مقدار دفعاتی که $\boldsymbol{v}$ در کلاس c_i رخ دهد را با $n_i(\boldsymbol{v})$ نشان می‌دهیم.

$$\hat{P}(c_i) = \frac{n_i}{n} \qquad \hat{f}(\mathbf{v}|c_i) = \frac{n_i(\mathbf{v})}{n_i}$$

برای بعضی از دسته‌ها هیچ مشاهده‌ای در نمونه وجود ندارد و برای پرهیز از شمارش صفر و تخصیص احتمال صفر به آن دسته‌ها پیشاپیش برای همه‌ی دسته‌ها یک مشاهده‌ی فرضی درنظر می‌گیریم.

$$\hat{f}(\mathbf{v}|c_i) = \frac{n_i(\mathbf{v}) + 1}{n_i + \prod_{j=1}^{d} m_j}$$

دسته‌بندی‌گر بیز(Bayes Classifier): مشخصه‌های دسته‌ای (Categorical Attributes)

Table 18.1. Discretized `sepal length` and `sepal width` attributes

Bins	Domain
[4.3, 5.2]	Very Short (a_{11})
(5.2, 6.1]	Short (a_{12})
(6.1, 7.0]	Long (a_{13})
(7.0, 7.9]	Very Long (a_{14})

(a) Discretized `sepal length`

Bins	Domain
[2.0, 2.8]	Short (a_{21})
(2.8, 3.6]	Medium (a_{22})
(3.6, 4.4]	Long (a_{23})

(b) Discretized `sepal width`

Table 18.2. Class-specific empirical (joint) probability mass function

	Class: c_1	X_2			$\hat{f}_{X_1}$
		Short ($\mathbf{e}_{21}$)	Medium ($\mathbf{e}_{22}$)	Long ($\mathbf{e}_{23}$)	
X_1	Very Short ($\mathbf{e}_{11}$)	1/50	33/50	5/50	39/50
	Short ($\mathbf{e}_{12}$)	0	3/50	8/50	11/50
	Long ($\mathbf{e}_{13}$)	0	0	0	0
	Very Long ($\mathbf{e}_{14}$)	0	0	0	0
	$\hat{f}_{X_2}$	1/50	36/50	13/50	

	Class: c_2	X_2			$\hat{f}_{X_1}$
		Short ($\mathbf{e}_{21}$)	Medium ($\mathbf{e}_{22}$)	Long ($\mathbf{e}_{23}$)	
X_1	Very Short ($\mathbf{e}_{11}$)	6/100	0	0	6/100
	Short ($\mathbf{e}_{12}$)	24/100	15/100	0	39/100
	Long ($\mathbf{e}_{13}$)	13/100	30/100	0	43/100
	Very Long ($\mathbf{e}_{14}$)	3/100	7/100	2/100	12/100
	$\hat{f}_{X_2}$	46/100	52/100	2/100	

$$\mathbf{x} = (5.3, 3.0)^T$$

$$\hat{P}(\mathbf{x}|c_1) = \hat{f}(\mathbf{v}|c_1) = 3/50 = 0.06$$
$$\hat{P}(\mathbf{x}|c_2) = \hat{f}(\mathbf{v}|c_2) = 15/100 = 0.15$$
$$\hat{P}(c_1|\mathbf{x}) \propto 0.06 \times 0.33 = 0.0198$$
$$\hat{P}(c_2|\mathbf{x}) \propto 0.15 \times 0.67 = 0.1005$$

$$\mathbf{x} = (6.75, 4.25)^T$$

$$\hat{P}(\mathbf{x}|c_1) = \hat{f}(\mathbf{v}|c_1) = \frac{0+1}{50+12} = 1.61 \times 10^{-2}$$
$$\hat{P}(\mathbf{x}|c_2) = \hat{f}(\mathbf{v}|c_2) = \frac{0+1}{100+12} = 8.93 \times 10^{-3}$$
$$\hat{P}(c_1|\mathbf{x}) \propto (1.61 \times 10^{-2}) \times 0.33 = 5.32 \times 10^{-3}$$
$$\hat{P}(c_2|\mathbf{x}) \propto (8.93 \times 10^{-3}) \times 0.67 = 5.98 \times 10^{-3}$$

مشخصه‌های عددی (Numerical Attributes)

برای سادگی متغیرهای تصادفی مستقل از هم در نظر گرفته می‌شوند.

$$P(\mathbf{x}|c_i) = P(x_1, x_2, \ldots, x_d|c_i) = \prod_{j=1}^{d} P(x_j|c_i)$$

$$|\mathbf{\Sigma}_i| = \det(\mathbf{\Sigma}_i) = \sigma_{i1}^2 \sigma_{i2}^2 \cdots \sigma_{id}^2 = \prod_{j=1}^{d} \sigma_{ij}^2$$

$$\mathbf{\Sigma}_i^{-1} = \begin{pmatrix} \frac{1}{\sigma_{i1}^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sigma_{i2}^2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \ldots & \frac{1}{\sigma_{id}^2} \end{pmatrix}$$

$$P(x_j|c_i) \propto f(x_j|\mu_{ij}, \sigma_{ij}^2) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}\right\}$$

Algorithm 18.2: Naive Bayes Classifier

NAIVEBAYES (D):

for $i = 1, \ldots, k$ **do**
  $\mathbf{D}_i \leftarrow \{\mathbf{x}_j^T \mid y_j = c_i, j = 1, \ldots, n\}$ // class-specific subsets
  $n_i \leftarrow |\mathbf{D}_i|$ // cardinality
  $\hat{P}(c_i) \leftarrow n_i/n$ // prior probability
  $\hat{\boldsymbol{\mu}}_i \leftarrow \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{D}_i} \mathbf{x}_j$ // mean
  $\overline{\mathbf{D}}_i = \mathbf{D}_i - \mathbf{1} \cdot \hat{\boldsymbol{\mu}}_i^T$ // centered data for class c_i
  for $j = 1, .., d$ **do** // class-specific var for jth attribute
    $\hat{\sigma}_{ij}^2 \leftarrow \frac{1}{n_i} (\overline{X}_j^i)^T (\overline{X}_j^i)$ // variance
  $\hat{\boldsymbol{\sigma}}_i \leftarrow \left(\hat{\sigma}_{i1}^2, \ldots, \hat{\sigma}_{id}^2\right)^T$ // class-specific attribute variances
return $\hat{P}(c_i), \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i$ for all $i = 1, \ldots, k$

TESTING (x and $\hat{P}(c_i)$, $\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\sigma}}_i$, for all $i \in [1, k]$):

$$\hat{y} \leftarrow \arg\max_{c_i} \left\{ \hat{P}(c_i) \prod_{j=1}^{d} f(x_j | \hat{\mu}_{ij}, \hat{\sigma}_{ij}^2) \right\}$$

return $\hat{y}$

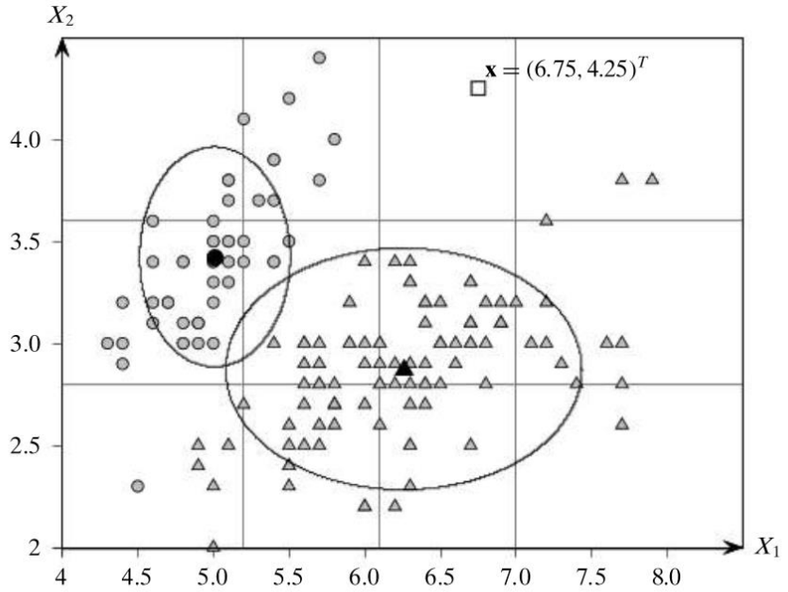


Figure 18.2. Naive Bayes: X_1:sepallength versus X_2:sepalwidth . The class means are shown in black; the density contours are also shown. The square represents a test point labeled **x**.

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.1218 & 0 \\ 0 & 0.1423 \end{pmatrix} \qquad \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.435 & 0 \\ 0 & 0.1096 \end{pmatrix}$$

$$\hat{P}(c_1|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)\hat{P}(c_1) = (4.014 \times 10^{-7}) \times 0.33 = 1.325 \times 10^{-7}$$

$$\hat{P}(c_2|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_2)\hat{P}(c_2) = (9.585 \times 10^{-5}) \times 0.67 = 6.422 \times 10^{-5}$$

$\hat{P}(c_2|\mathbf{x}) > \hat{P}(c_1|\mathbf{x})$ the class for **x** is predicted as $\hat{y} = c_2$.

Table 18.1. Discretized `sepal length` and `sepal width` attributes

Bins	Domain
[4.3, 5.2]	Very Short (a_{11})
(5.2, 6.1]	Short (a_{12})
(6.1, 7.0]	Long (a_{13})
(7.0, 7.9]	Very Long (a_{14})

(a) Discretized `sepal length`

Bins	Domain
[2.0, 2.8]	Short (a_{21})
(2.8, 3.6]	Medium (a_{22})
(3.6, 4.4]	Long (a_{23})

(b) Discretized `sepal width`

Table 18.2. Class-specific empirical (joint) probability mass function

	Class: c_1	X_2			$\hat{f}_{X_1}$
		Short ($\mathbf{e}_{21}$)	Medium ($\mathbf{e}_{22}$)	Long ($\mathbf{e}_{23}$)	
X_1	Very Short ($\mathbf{e}_{11}$)	1/50	33/50	5/50	39/50
	Short ($\mathbf{e}_{12}$)	0	3/50	8/50	11/50
	Long ($\mathbf{e}_{13}$)	0	0	0	0
	Very Long ($\mathbf{e}_{14}$)	0	0	0	0
	$\hat{f}_{X_2}$	1/50	36/50	13/50	

	Class: c_2	X_2			$\hat{f}_{X_1}$
		Short ($\mathbf{e}_{21}$)	Medium ($\mathbf{e}_{22}$)	Long ($\mathbf{e}_{23}$)	
X_1	Very Short ($\mathbf{e}_{11}$)	6/100	0	0	6/100
	Short ($\mathbf{e}_{12}$)	24/100	15/100	0	39/100
	Long ($\mathbf{e}_{13}$)	13/100	30/100	0	43/100
	Very Long ($\mathbf{e}_{14}$)	3/100	7/100	2/100	12/100
	$\hat{f}_{X_2}$	46/100	52/100	2/100	

مشخصه‌های دسته‌ای (Categorical Attributes)

$$P(\mathbf{x}|c_i) = \prod_{j=1}^{d} P(x_j|c_i) = \prod_{j=1}^{d} f(\mathbf{X}_j = \mathbf{e}_{jr_j} \mid c_i)$$

$$\hat{f}(\mathbf{v}_j|c_i) = \frac{n_i(\mathbf{v}_j)}{n_i}$$

شبهه شمارش (Pseudo-Counting) در صورت عدم مشاهده در نمونه

$$\hat{f}(\mathbf{v}_j|c_i) = \frac{n_i(\mathbf{v}_j)+1}{n_i+m_j}$$

$\mathbf{x} = (6.75, 4.25)$

$$\hat{P}(\mathbf{v}|c_1) = \hat{P}(\mathbf{e}_{13}|c_1) \cdot \hat{P}(\mathbf{e}_{23}|c_1) = \left(\frac{0+1}{50+4}\right) \cdot \left(\frac{13}{50}\right) = 4.81 \times 10^{-3}$$

$$\hat{P}(\mathbf{v}|c_2) = \hat{P}(\mathbf{e}_{13}|c_2) \cdot \hat{P}(\mathbf{e}_{23}|c_2) = \left(\frac{43}{100}\right) \cdot \left(\frac{2}{100}\right) = 8.60 \times 10^{-3}$$

$$\hat{P}(c_1|\mathbf{v}) \propto (4.81 \times 10^{-3}) \times 0.33 = 1.59 \times 10^{-3}$$

$$\hat{P}(c_2|\mathbf{v}) \propto (8.6 \times 10^{-3}) \times 0.67 = 5.76 \times 10^{-3}$$

دسته‌بندی‌گر K نزدیکترین همسایگان (K Nearest Neighbors Classifier)

در بخش‌های قبل یک رویکرد پارامتری برای تخمین درستنمایی $P(x|c_i)$ در نظر گرفتیم و اینک مانند روش برآورد چگالی (Density Estimation)، یک رویکرد ناپارامتریک را در نظر می‌گیریم، که هیچ فرضی در مورد تابع چگالی احتمال مشترک ندارد و مستقیما از داده‌ها تابع چگالی را تخمین می‌زند.

مجموعه داده‌ی یادگیری $\boldsymbol{D} \in \mathbb{R}^{n \times d}$ را در نظر بگیرید و $\boldsymbol{D}_i \subseteq \boldsymbol{D}$ داده‌هایی از آن مجموعه می‌باشند که برچسب دسته‌ی c_i دارند و به آن دسته نسبت داده می‌شوند.

فاصله‌ی نقطه‌ی آزمون $\boldsymbol{x} \in \mathbb{R}^d$ تا Kامین نزدیکترین همسایه در مجموعه $\boldsymbol{D}$ را r بگیرید.
پس در کره‌ای به شعاع r حول $\boldsymbol{x}$ تعداد K نقطه وجود دارد که تعداد K_i تای آن‌ها برچسب c_i دارند.

$$B_d(\mathbf{x}, r) = \{\mathbf{x}_i \in \mathbf{D} \mid \|\mathbf{x} - \mathbf{x}_i\| \leq r\}$$

$$|B_d(\mathbf{x}, r)| = K. \qquad K_i = \{\mathbf{x}_j \in B_d(\mathbf{x}, r) \mid y_j = c_i\}$$

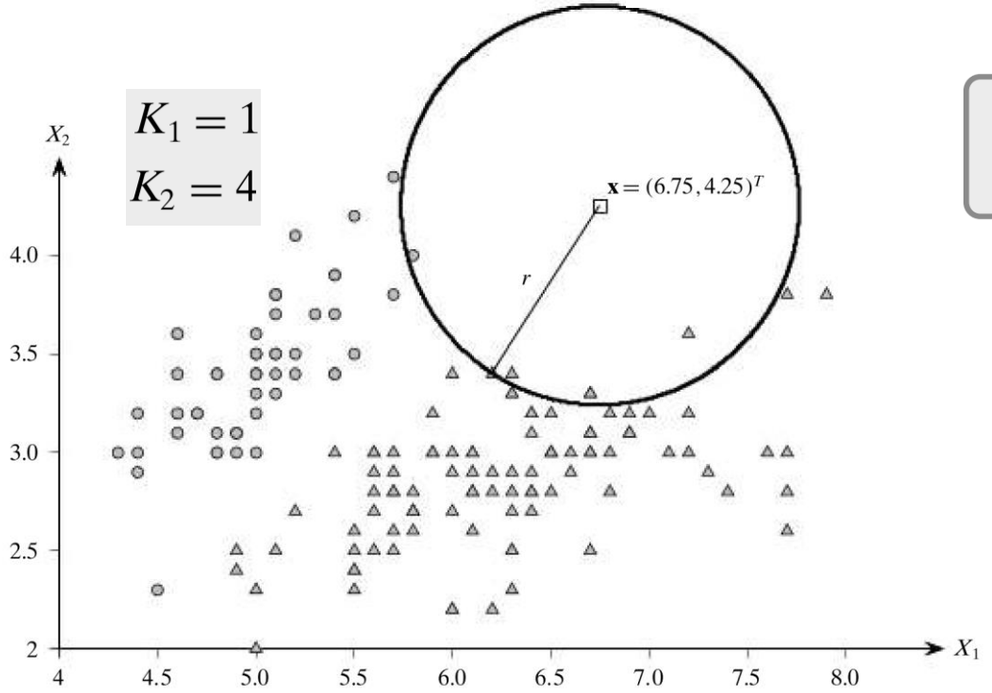


Figure 18.3. Iris Data: K Nearest Neighbors Classifier

$$\hat{f}(\mathbf{x}|c_i) = \frac{K_i/n_i}{V} = \frac{K_i}{n_i V}$$

$$\hat{f}(\mathbf{x}|c_i)\hat{P}(c_i) = \frac{K_i}{n_i V} \cdot \frac{n_i}{n} = \frac{K_i}{nV}$$

$$V = \text{vol}(B_d(\mathbf{x}, r))$$

$$P(c_i|\mathbf{x}) = \frac{\hat{f}(\mathbf{x}|c_i)\hat{P}(c_i)}{\sum_{j=1}^{k} \hat{f}(\mathbf{x}|c_j)\hat{P}(c_j)}$$

$$P(c_i|\mathbf{x}) = \frac{\frac{K_i}{nV}}{\sum_{j=1}^{k} \frac{K_j}{nV}} = \frac{K_i}{K}$$

$$\hat{y} = \arg\max_{c_i}\{P(c_i|\mathbf{x})\} = \arg\max_{c_i}\left\{\frac{K_i}{K}\right\} = \arg\max_{c_i}\{K_i\}$$