Wrocław University
of Science and Technology

Modele generujące
GDA, Naive Bayes

Szymon Zaręba

szymon.zareba@pwr.edu.pl

Modelowanie generujące

Rozkład warunkowy $p(y|\mathbf{x})$ można wyznaczyć korzystając ze wzoru Bayesa i twierdzenia o prawdopodobieństwie całkowitym

$$\begin{aligned} p(y|\mathbf{x}) &= \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|y)p(y)}{\sum_{y'} p(\mathbf{x}|y')p(y')} \end{aligned}$$

Wielkości, które będą modelowane to $p(\mathbf{x}|y, \mathbf{\Theta})$ oraz $p(y|\mathbf{\Theta})$.

Podejście generujące: Naive Bayes

Rozkład $p(y|\mathbf{\Theta})$ będzie rozkładem wielopunktowym:

$$p(y|\mathbf{\Theta}) = \mathrm{M}(y|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{y^{(k)}}$$

Model Naive Bayes zakłada niezależność cech:

$$p(\mathbf{x}|y, \mathbf{\Theta}) = \prod_{d=1}^{D} p(x^{(d)}|y, \mathbf{\Theta})$$

Podejście generujące: Naive Bayes

Przypadek dyskretny (1)

Każda z cech modelowana jest rozkładem dwupunktowym:

$$p(x^{(d)}|y = k, \mathbf{\Theta}) = \mathrm{B}(x^{(d)}|\theta_{k,d}) = \theta_{k,d}^{x^{(d)}}(1 - \theta_{k,d})^{1-x^{(d)}}$$

Rozkład $p(\mathbf{x}|y = k, \mathbf{\Theta})$ dla wektora x przyjmuje postać:

$$p(\mathbf{x}|y = k, \mathbf{\Theta}) = \prod_{d=1}^{D} \mathrm{B}(x^{(d)}|\theta_{k,d})$$

Podejście generujące: Naive Bayes

Przypadek dyskretny (2)

W ogólnym przypadku rozkład $p(\mathbf{x}|y, \mathbf{\Theta})$ przyjmuje postać:

$$p(\mathbf{x}|y, \mathbf{\Theta}) = \prod_{k=1}^{K} \left[\prod_{d=1}^{D} \mathrm{B}(x^{(d)}|\theta_{k,d}) \right]^{y^{(k)}}$$

Zatem rozkład łączny $p(\mathbf{x}, y|\mathbf{\Theta})$ można zapisać jako:

$$p(\mathbf{x}, y|\mathbf{\Theta}) = \prod_{k=1}^{K} \left[\pi_k \prod_{d=1}^{D} \mathrm{B}(x^{(d)}|\theta_{k,d}) \right]^{y^{(k)}}$$

Podejście generujące: Naive Bayes

Przypadek ciągły (1)

Każda z cech modelowana jest rozkładem normalnym:

$$p(x^{(d)}|y=k,\mathbf{\Theta}) = \mathcal{N}(x^{(d)}|\mu_{k,d},\sigma^2_{k,d})$$

Rozkład $p(\mathbf{x}|y=k,\mathbf{\Theta})$ dla wektora x przyjmuje postać:

$$p(\mathbf{x}|y=k,\mathbf{\Theta}) = \prod_{d=1}^{D}\mathcal{N}(x^{(d)}|\mu_{k,d},\sigma^2_{k,d}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)$$

$$\boldsymbol{\mu}_k = \begin{bmatrix}\mu_{k,1}\\ \cdots \\ \mu_{k,D}\end{bmatrix} \qquad \mathbf{\Sigma}_k = \mathrm{diag}(\sigma^2_k) = \begin{bmatrix}\sigma^2_{k,1} & \cdots & 0\\ \cdots & \cdots & \cdots\\ 0 & \cdots & \sigma^2_{k,D}\end{bmatrix}$$

Podejście generujące: Naive Bayes

Przypadek ciągły (2)

W ogólnym przypadku rozkład $p(\mathbf{x}|y, \mathbf{\Theta})$ przyjmuje postać:

$$p(\mathbf{x}|y, \mathbf{\Theta}) = \prod_{k=1}^{K} \left[\prod_{d=1}^{D} \mathcal{N}(x^{(d)}|\mu_{k,d}, \sigma^2_{k,d}) \right]^{y^{(k)}}$$

Zatem rozkład łączny $p(\mathbf{x}, y|\mathbf{\Theta})$ można zapisać jako:

$$p(\mathbf{x}, y|\mathbf{\Theta}) = \prod_{k=1}^{K} \left[\pi_k \prod_{d=1}^{D} \mathcal{N}(x^{(d)}|\mu_{k,d}, \sigma^2_{k,d}) \right]^{y^{(k)}}$$

Podejście generujące: GDA

Można zadać następujące rozkłady $p(y|\mathbf{\Theta})$ i $p(\mathbf{x}|y=k,\mathbf{\Theta})$:

$$p(y|\mathbf{\Theta}) = \mathrm{M}(y|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{y^{(k)}}$$

$$p(\mathbf{x}|y=k,\mathbf{\Theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)$$

W ogólnym przypadku rozkład $p(\mathbf{x}|y,\mathbf{\Theta})$ przyjmuje postać:

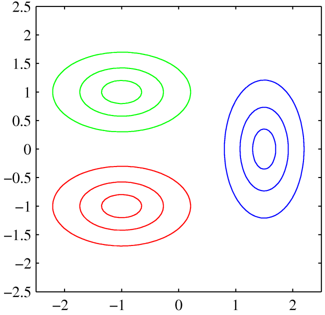
$$p(\mathbf{x}|y,\mathbf{\Theta}) = \prod_{k=1}^{K} \left[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)\right]^{y^{(k)}}$$

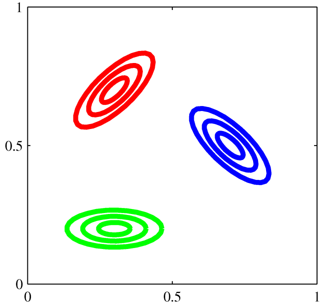
Zatem rozkład łączny $p(\mathbf{x},y|\mathbf{\Theta})$ można zapisać jako:

$$p(\mathbf{x},y|\mathbf{\Theta}) = \prod_{k=1}^{K} \left[\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)\right]^{y^{(k)}}$$

Zestawienie (1)

Model	$p(x, y)$	$\boldsymbol{\Theta}$	$\mathbf{card}(\boldsymbol{\Theta})$
NB-d	$\prod_{k=1}^{K}\left[\pi_k \prod_{d=1}^{D} \mathrm{B}(x^{(d)} \mid \theta_{d,k})\right]^{y^{(k)}}$	$\pi_1, \cdots, \pi_K$, $\theta_{1,1}, \cdots, \theta_{D,K}$	K KD
NB-c	$\prod_{k=1}^{K}\left[\pi_k \prod_{d=1}^{D} \mathcal{N}(x^{(d)} \mid \mu_{k,d}, \sigma^2_{k,d})\right]^{y^{(k)}}$	$\pi_1, \cdots, \pi_K$, $\mu_{1,1}, \cdots, \mu_{D,K}$, $\sigma^2_{1,1}, \cdots, \sigma^2_{D,K}$	K KD KD
GDA	$\prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{y^{(k)}}$	$\pi_1, \cdots, \pi_K$, $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K$, $\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K$	K KD $K\frac{D(D+1)}{2}$





Uczenie

GDA

Dany jest rozkład $p(\mathbf{x}|y, \mathbf{\Theta})$:

$$p(\mathbf{x}|y, \mathbf{\Theta}) = \prod_{k=1}^{K} \left[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right]^{y^{(k)}}$$

Można wyznaczyć funkcję wiarygodności:

Uczenie

GDA

Dany jest rozkład $p(\mathbf{x}|y, \mathbf{\Theta})$:

$$p(\mathbf{x}|y, \mathbf{\Theta}) = \prod_{k=1}^{K} \left[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right]^{y^{(k)}}$$

Można wyznaczyć funkcję wiarygodności:

$$p(\mathcal{D}|\mathbf{\Theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right]^{y_n^{(k)}}$$

Uczenie

GDA

Dany jest rozkład $p(\mathbf{x}|y, \mathbf{\Theta})$:

$$p(\mathbf{x}|y, \mathbf{\Theta}) = \prod_{k=1}^{K} \left[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right]^{y^{(k)}}$$

Można wyznaczyć funkcję wiarygodności:

$$p(\mathcal{D}|\mathbf{\Theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right]^{y_n^{(k)}}$$

po wstawieniu rozkładu normalnego:

Uczenie

GDA

Dany jest rozkład $p(\mathbf{x}|y, \mathbf{\Theta})$:

$$p(\mathbf{x}|y, \mathbf{\Theta}) = \prod_{k=1}^{K} \left[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right]^{y^{(k)}}$$

Można wyznaczyć funkcję wiarygodności:

$$p(\mathcal{D}|\mathbf{\Theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right]^{y_n^{(k)}}$$

po wstawieniu rozkładu normalnego:

$$p(\mathcal{D}|\mathbf{\Theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T \mathbf{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\right\}\right]^{y_n^{(k)}}$$

Uczenie
GDA

Można wyznaczyć logarytm funkcji wiarygodności:

Uczenie
GDA

Można wyznaczyć logarytm funkcji wiarygodności:

$$\ln p(\mathcal{D}|\mathbf{\Theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K} y_n^{(k)} \left[-\frac{D}{2}\ln\left(2\pi\right) - \frac{1}{2}\ln\left|\mathbf{\Sigma}_k\right| - \frac{1}{2}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T \mathbf{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\right]$$

Uczenie

GDA

Można wyznaczyć logarytm funkcji wiarygodności:

$$\ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N} \sum_{k=1}^{K} y_n^{(k)} \left[-\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln|\boldsymbol{\Sigma}_k| - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right]$$

wyznaczyć jego gradient ze względu na parametr $\boldsymbol{\mu}_j$ i $\boldsymbol{\Sigma}_j$:

Uczenie

GDA

Można wyznaczyć logarytm funkcji wiarygodności:

$$\ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K} y_n^{(k)} \left[-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right]$$

wyznaczyć jego gradient ze względu na parametr $\boldsymbol{\mu}_j$ i $\boldsymbol{\Sigma}_j$:

$$\nabla_{\boldsymbol{\mu}_j} \ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N} y_n^{(j)} \boldsymbol{\Sigma}_j^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)$$

$$\nabla_{\boldsymbol{\Sigma}_j} \ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N} y_n^{(j)} \left[-\frac{1}{2}\boldsymbol{\Sigma}_j^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_j^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1}\right]$$

Uczenie

GDA

Dla parametru $\boldsymbol{\mu}_j$:

$$\nabla_{\boldsymbol{\mu}_j} \ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N} y_n^{(j)} \boldsymbol{\Sigma}_j^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)$$

Uczenie

GDA

Dla parametru $\boldsymbol{\mu}_j$:

$$\nabla_{\boldsymbol{\mu}_j} \ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N} y_n^{(j)} \boldsymbol{\Sigma}_j^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)$$

$$\boldsymbol{\mu}_j^{\mathsf{ML}} = \frac{1}{N_j} \sum_{n=1}^{N} y_n^{(j)} \mathbf{x}_n$$

Uczenie

GDA

Dla parametru $\boldsymbol{\mu}_j$:

$$\nabla_{\boldsymbol{\mu}_j} \ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N} y_n^{(j)} \boldsymbol{\Sigma}_j^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)$$

$$\boldsymbol{\mu}_j^{\mathsf{ML}} = \frac{1}{N_j} \sum_{n=1}^{N} y_n^{(j)} \mathbf{x}_n$$

Dla parametru $\boldsymbol{\Sigma}_j$:

$$\nabla_{\boldsymbol{\Sigma}_j} \ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N} y_n^{(j)} \left[-\frac{1}{2}\boldsymbol{\Sigma}_j^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_j^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1}\right]$$

Uczenie

GDA

Dla parametru $\boldsymbol{\mu}_j$:

$$\nabla_{\boldsymbol{\mu}_j} \ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N} y_n^{(j)} \boldsymbol{\Sigma}_j^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)$$

$$\boldsymbol{\mu}_j^{\mathsf{ML}} = \frac{1}{N_j} \sum_{n=1}^{N} y_n^{(j)} \mathbf{x}_n$$

Dla parametru $\boldsymbol{\Sigma}_j$:

$$\nabla_{\boldsymbol{\Sigma}_j} \ln p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{n=1}^{N} y_n^{(j)} \left[-\frac{1}{2}\boldsymbol{\Sigma}_j^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_j^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1}\right]$$

$$\boldsymbol{\Sigma}_j^{\mathsf{ML}} = \frac{1}{N_j} \sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)^T$$

Zestawienie (2)

Model	Estymatory
NB-d	$\pi_j = \frac{N_j}{N}$ $\theta_{j,d} = \frac{1}{N_j} \sum_{n=1}^{N} y_n^{(j)} x_n^{(d)}$
NB-c	$\pi_j = \frac{N_j}{N}$ $\boldsymbol{\mu}_{j,d} = \frac{1}{N_j} \sum_{n=1}^{N} y_n^{(j)} x_n^{(d)}$ $\boldsymbol{\sigma}_{j,d} = \frac{1}{N_j} \sum_{n=1}^{N} \left(x_n^{(d)} - \mu_{j,d} \right)^2$
GDA	$\pi_j = \frac{N_j}{N}$ $\boldsymbol{\mu}_j = \frac{1}{N_j} \sum_{n=1}^{N} y_n^{(j)} \mathbf{x}_n$ $\boldsymbol{\Sigma}_j = \frac{1}{N_j} \sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}_j \right) \left(\mathbf{x}_n - \boldsymbol{\mu}_j \right)^T$

Predykcja

GDA

Wykorzystując wzór Bayesa i twierdzenie o prawdopodobieństwie całkowitym:

$$p(y=j|\mathbf{x}) = \frac{p(\mathbf{x}, y=j)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y=j)p(y=j)}{\sum\limits_{k=1}^{K} p(\mathbf{x}, y=k)} = \frac{p(\mathbf{x}|y=j)p(y=j)}{\sum\limits_{k=1}^{K} p(\mathbf{x}|y=k)p(y=k)}$$

Dla modelu GDA:

$$p(y=j|\mathbf{x}) = \frac{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j}{\sum\limits_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}$$

Zestawienie (3)

Model	Predykcja
NB-d	$\dfrac{\pi_j \prod_{d=1}^{D} \theta_{j,d}^{x^{(d)}} \left(1 - \theta_{j,d}\right)^{1-x^{(d)}}}{\sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \theta_{k,d}^{x^{(d)}} \left(1 - \theta_{k,d}\right)^{1-x^{(d)}}}$
NB-c	$\dfrac{\pi_j \prod_{d=1}^{D} \mathcal{N}\left(x^{(d)} \mid \mu_{j,d}, \sigma_{j,d}^2\right)}{\sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \mathcal{N}\left(x^{(d)} \mid \mu_{k,d}, \sigma_{k,d}^2\right)}$
GDA	$\dfrac{\pi_j \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$

Predykcja

Teoria decyzji mówi, że optymalnym wyborem jest y, który minimalizuje ryzyko (średnią stratę) przy zadanej funkcji straty. Mając:

- $p(y|x, \boldsymbol{\Theta})$ - model
- $y \in \{1, \cdots, K\}$ - zbiór możliwych decyzji
- $L(y, \hat{y}) = \begin{cases} 1, y \neq \hat{y} \\ 0, y = \hat{y} \end{cases}$ - funkcja straty

Predykcja

Teoria decyzji mówi, że optymalnym wyborem jest y, który minimalizuje ryzyko (średnią stratę) przy zadanej funkcji straty. Mając:

- $p(y|x, \mathbf{\Theta})$ - model
- $y \in \{1, \cdots, K\}$ - zbiór możliwych decyzji
- $L(y, \hat{y}) = \begin{cases} 1, y \neq \hat{y} \\ 0, y = \hat{y} \end{cases}$ - funkcja straty

$R(\hat{y}) = \mathbb{E}[L(y, \hat{y})]$

Predykcja

Teoria decyzji mówi, że optymalnym wyborem jest y, który minimalizuje ryzyko (średnią stratę) przy zadanej funkcji straty. Mając:

- $p(y|x, \mathbf{\Theta})$ - model
- $y \in \{1, \cdots, K\}$ - zbiór możliwych decyzji
- $L(y, \hat{y}) = \begin{cases} 1, y \neq \hat{y} \\ 0, y = \hat{y} \end{cases}$ - funkcja straty

$$R(\hat{y}) = \mathbb{E}[L(y, \hat{y})] = \sum_{k=1}^{K} L(y = k, \hat{y}) p(y = k|\mathbf{x})$$

Predykcja

Teoria decyzji mówi, że optymalnym wyborem jest y, który minimalizuje ryzyko (średnią stratę) przy zadanej funkcji straty. Mając:

- $p(y|x, \mathbf{\Theta})$ - model
- $y \in \{1, \cdots, K\}$ - zbiór możliwych decyzji
- $L(y, \hat{y}) = \begin{cases} 1, y \neq \hat{y} \\ 0, y = \hat{y} \end{cases}$ - funkcja straty

$$R(\hat{y}) = \mathbb{E}[L(y, \hat{y})] = \sum_{k=1}^{K} L(y = k, \hat{y}) p(y = k|\mathbf{x})$$
$$= L(y = 1, \hat{y}) p(y = 1|\mathbf{x}) + \cdots + L(y = K, \hat{y}) p(y = K|\mathbf{x})$$

Predykcja

Teoria decyzji mówi, że optymalnym wyborem jest y, który minimalizuje ryzyko (średnią stratę) przy zadanej funkcji straty. Mając:

- $p(y|x, \mathbf{\Theta})$ - model
- $y \in \{1, \cdots, K\}$ - zbiór możliwych decyzji
- $L(y, \hat{y}) = \begin{cases} 1, y \neq \hat{y} \\ 0, y = \hat{y} \end{cases}$ - funkcja straty

$$
\begin{aligned}
R(\hat{y}) = \mathbb{E}[L(y, \hat{y})] &= \sum_{k=1}^{K} L(y = k, \hat{y}) p(y = k|\mathbf{x}) \\
&= L(y = 1, \hat{y}) p(y = 1|\mathbf{x}) + \cdots + L(y = K, \hat{y}) p(y = K|\mathbf{x})
\end{aligned}
$$

$$
\hat{y} = \operatorname*{argmax}_{k} p(y = k|\mathbf{x})
$$