MAP Estimation and Bayesian

Iran University of Science and Technology

M. S. Tahaei PhD.

Fall 2024

Courtesy: slides are adopted partly from Dr. Soleymani, Sharif University

Outline

- Maximum A Posteriori (MAP) estimation
- Bayes classifier
- Naïve Bayes classifier

Maximum A Posteriori (MAP) estimation

- MAP estimation

$$\boldsymbol{\theta}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\mathcal{D})$$

- Since $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

$$\boldsymbol{\theta}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- Example of prior distribution:

$$p(\theta) = \mathcal{N}(\theta_0, \sigma^2)$$

MAP estimation Gaussian: unknown $\boldsymbol{\mu}$

$p(x|\mu) \sim N(\mu, \sigma^2)$ — μ is the only unknown parameter

$p(\mu|\mu_0) \sim N(\mu_0, \sigma_0^2)$ — μ_0 and σ_0 are known

$$\frac{d}{d\mu} \ln\left(p(\mu) \prod_{1=i}^{N} p\left(x^{(i)} | \mu \right) \right) = 0$$

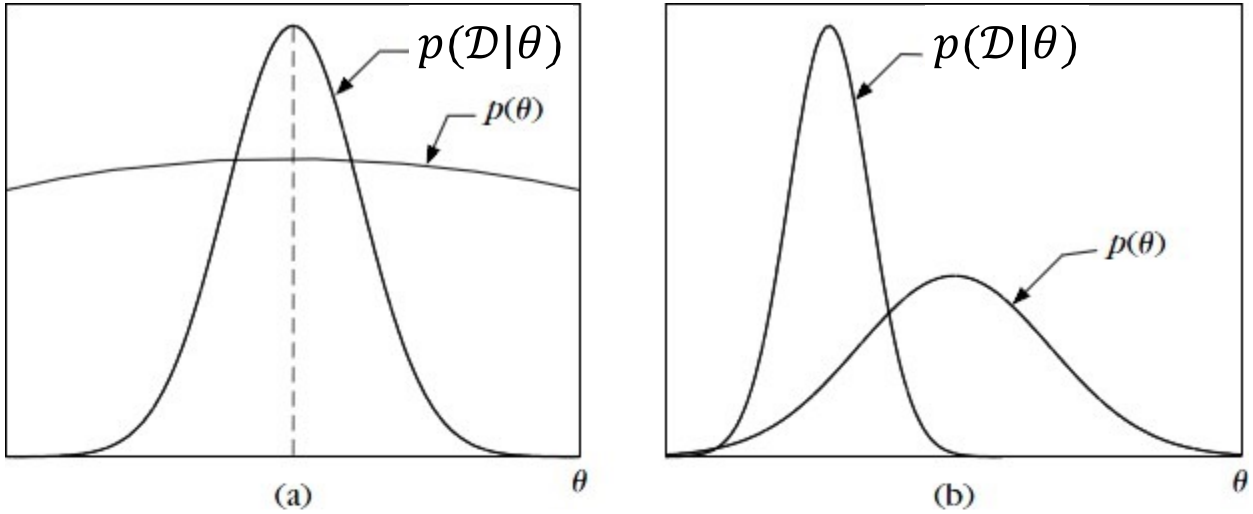
$$\Rightarrow \sum_{i=1}^{N} \frac{1}{\sigma^2} \left(x^{(i)} - \mu \right) - \frac{1}{\sigma_0^2} (\mu - \mu_0) = 0$$

$$\Rightarrow \boxed{\mu_{MAP} = \frac{\mu_0 + \frac{\sigma_0^2}{\sigma^2} \sum_{i=1}^{N} x^{(i)}}{1 + \frac{\sigma_0^2}{\sigma^2} N}}$$

$$\frac{\sigma_0^2}{\sigma^2} \gg 1 \text{ or } N \to \infty \Rightarrow \mu_{MAP} = \mu_{ML} = \frac{\sum_{i=1}^{N} x^{(i)}}{N}$$

Maximum A Posteriori (MAP) estimation

- Given a set of observations $\mathcal{D}$ and a prior distribution $p(\boldsymbol{\theta})$ on parameters, the parameter vector that maximizes $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ is found.

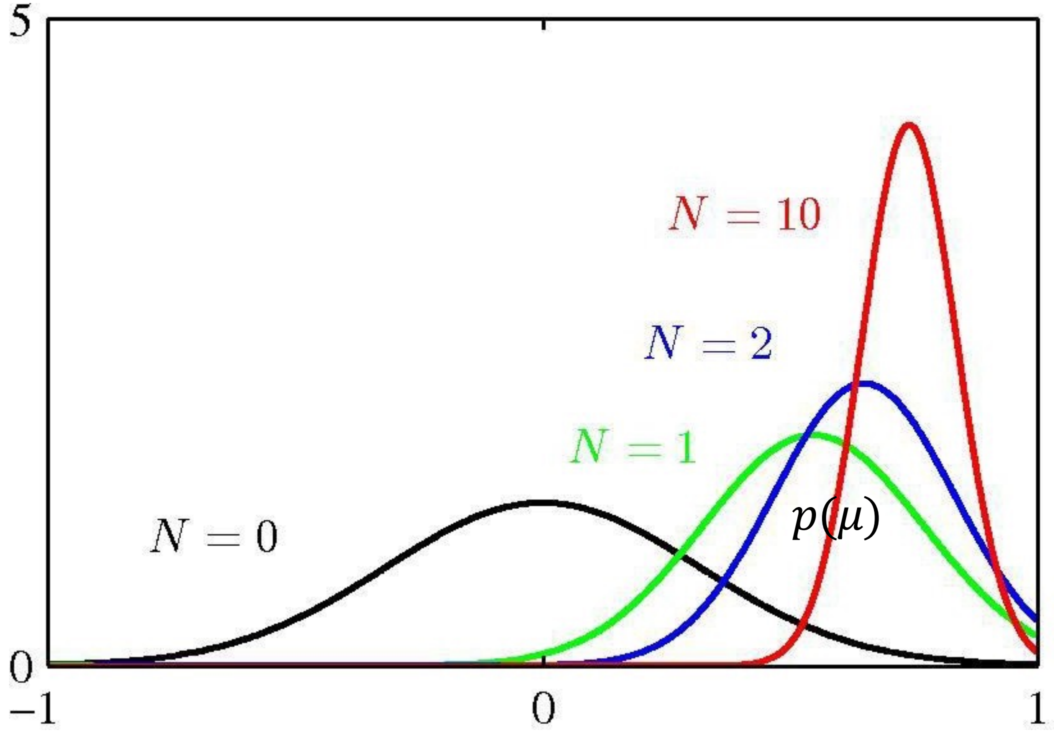


$\theta_{MAP} \cong \theta_{ML}$

$\theta_{MAP} > \theta_{ML}$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$$

MAP estimation
Gaussian: unknown $\boldsymbol{\mu}$ (known $\boldsymbol{\sigma}$)



[Bishop]

$$p(\mu|\mathcal{D}) \propto p(\mu)p(\mathcal{D}|\mu)$$

$$p(\mu|\mathcal{D}) = N(\mu|\mu_N, \sigma_N)$$

$$\mu_N = \frac{\mu_0 + \frac{\sigma_0^2}{\sigma^2}\sum_{i=1}^{N} x^{(i)}}{1 + \frac{\sigma_0^2}{\sigma^2} N}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

More samples $\Longrightarrow$ sharper $p(\mu|\mathcal{D})$
Higher confidence in estimation

Definitions

- Posterior probability: $p(\mathcal{C}_k|\boldsymbol{x})$
- Likelihood or class conditional probability: $p(\boldsymbol{x}|\mathcal{C}_k)$
- Prior probability: $p(\mathcal{C}_k)$

$p(\boldsymbol{x})$: pdf of feature vector $\boldsymbol{x}$ ($p(\boldsymbol{x}) = \sum_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{x})\, p(\mathcal{C}_k)$

$p(\boldsymbol{x}|\mathcal{C}_k)$: pdf of feature vector $\boldsymbol{x}$ for samples of class $\mathcal{C}_k$

$p(\mathcal{C}_k)$: probability of the label be $\mathcal{C}_k$

Bayes decision rule

If $P(\mathcal{C}_1|\boldsymbol{x}) > P(\mathcal{C}_2|\boldsymbol{x})$ decide $\mathcal{C}_1$
otherwise decide $\mathcal{C}_2$

$K = 2$

$$p(error|\boldsymbol{x}) = \begin{cases} p(C_2|\boldsymbol{x}) & \text{if we decide } \mathcal{C}_1 \\ P(C_1|\boldsymbol{x}) & \text{if we decide } \mathcal{C}_2 \end{cases}$$

- If we use Bayes decision rule:

$$P(error|\boldsymbol{x}) = \min\{P(\mathcal{C}_1|\boldsymbol{x}), P(\mathcal{C}_2|\boldsymbol{x})\}$$

Using Bayes rule, for each $\boldsymbol{x}$, $P(error|\boldsymbol{x})$ is as small as possible and thus this rule minimizes the probability of error

Optimal classifier

- The optimal decision is the one that minimizes the expected number of mistakes

- We show that Bayes classifier is an optimal classifier

Bayes theorem

- Bayes' theorem

Posterior

Likelihood

Prior

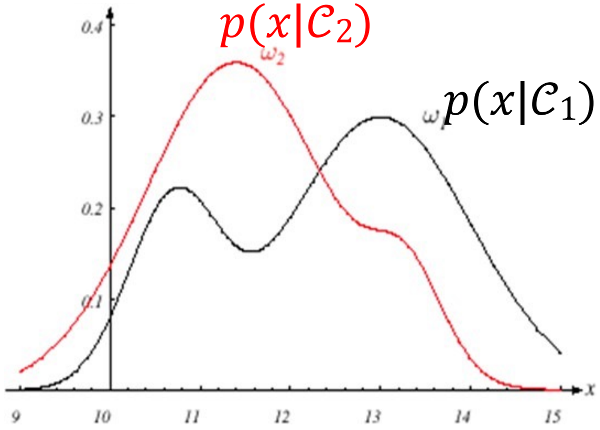
$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\boldsymbol{x})}$$

- Posterior probability: $p\,(\mathcal{C}_k|\boldsymbol{x})$
- Likelihood or class conditional probability: $p\,(\boldsymbol{x}|\mathcal{C}_k)$
- Prior probability: $p(\mathcal{C}_k)$

$p(\boldsymbol{x})$: pdf of feature vector $\boldsymbol{x}$ ($p\,(\boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)$)
$p(\boldsymbol{x}|\mathcal{C}_k)$: pdf of feature vector $\boldsymbol{x}$ for samples of class $\mathcal{C}_k$
$p(\mathcal{C}_k)$:probability of the label be $\mathcal{C}_k$

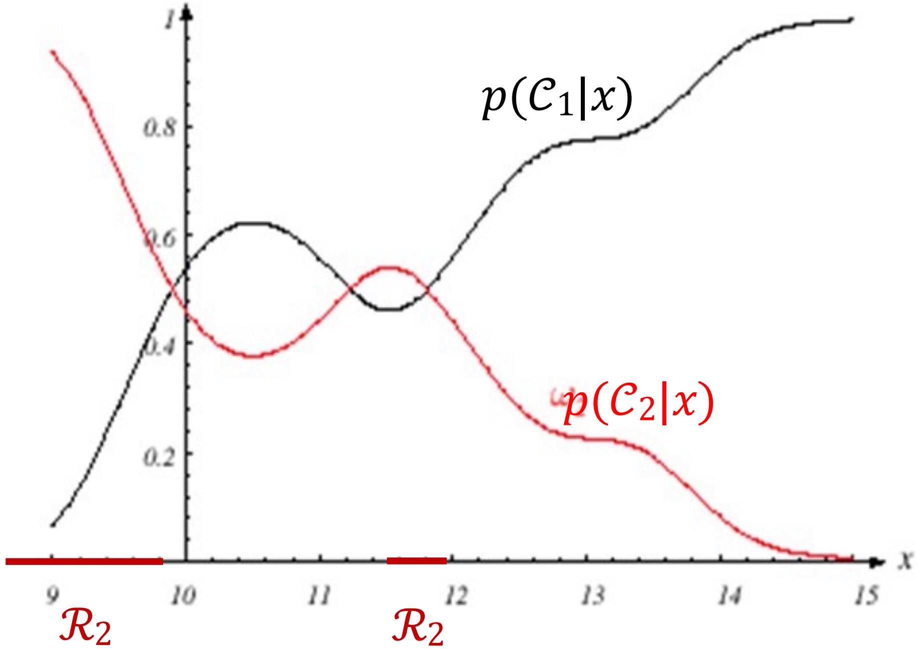
Bayes decision rule: example

- Bayes decision: Choose the class with highest $p\ (\mathcal{C}_k|\boldsymbol{x})$



$$p(\mathcal{C}_1) = \frac{2}{3}$$

$$p(\mathcal{C}_2) = \frac{1}{3}$$



$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x}) = p(\mathcal{C}_1)p(\boldsymbol{x}|\mathcal{C}_1) + p(\mathcal{C}_2)p(\boldsymbol{x}|\mathcal{C}_2)$$

Bayesian decision rule

- If $P(\mathcal{C}_1|\boldsymbol{x}) > P(\mathcal{C}_2|\boldsymbol{x})$ decide $\mathcal{C}_1$

  otherwise decide $\mathcal{C}_2$

Equivalent

- If $\frac{p(\boldsymbol{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{p(\boldsymbol{x})} > \frac{p(\boldsymbol{x}|\mathcal{C}_2)P(\mathcal{C}_2)}{p(x)}$ decide $\mathcal{C}_1$

  otherwise decide $\mathcal{C}_2$

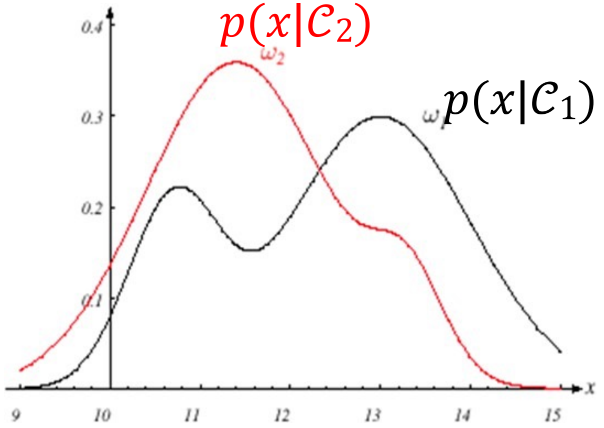
Equivalent

- If $p(\boldsymbol{x}|\mathcal{C}_1)P(\mathcal{C}_1) > p(\boldsymbol{x}|\mathcal{C}_2)P(\mathcal{C}_2)$ decide $\mathcal{C}_1$

  otherwise decide $\mathcal{C}_2$

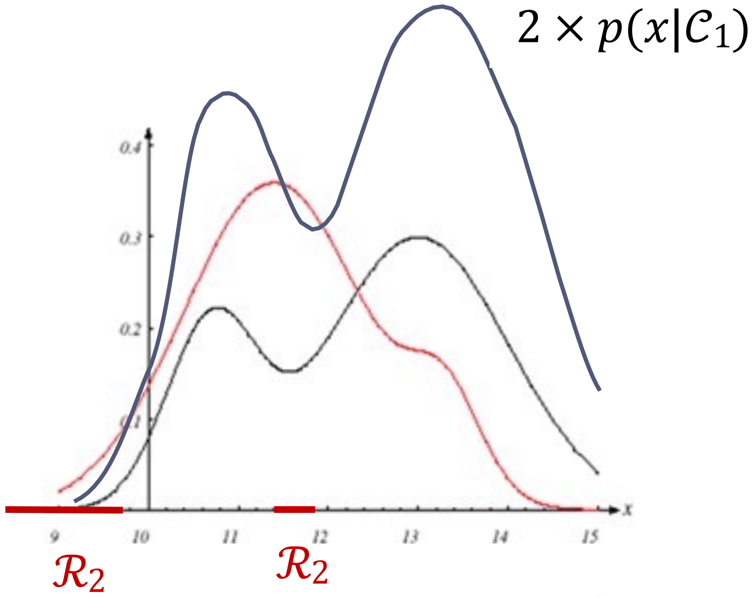
Bayes decision rule: example

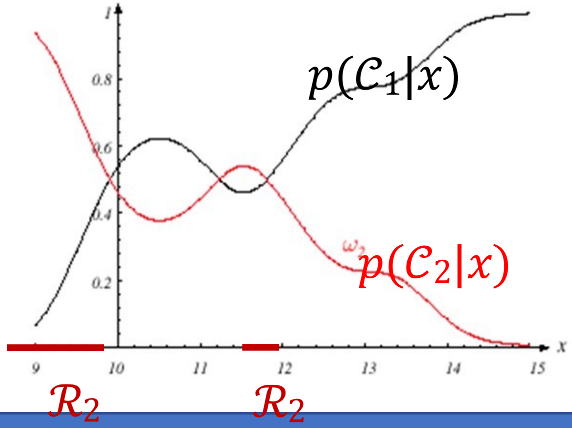
- Bayes decision: Choose the class with highest $p\ (\mathcal{C}_k|\boldsymbol{x})$



$$p(\mathcal{C}_1) = \frac{2}{3}$$

$$p(\mathcal{C}_2) = \frac{1}{3}$$





Bayes Classier

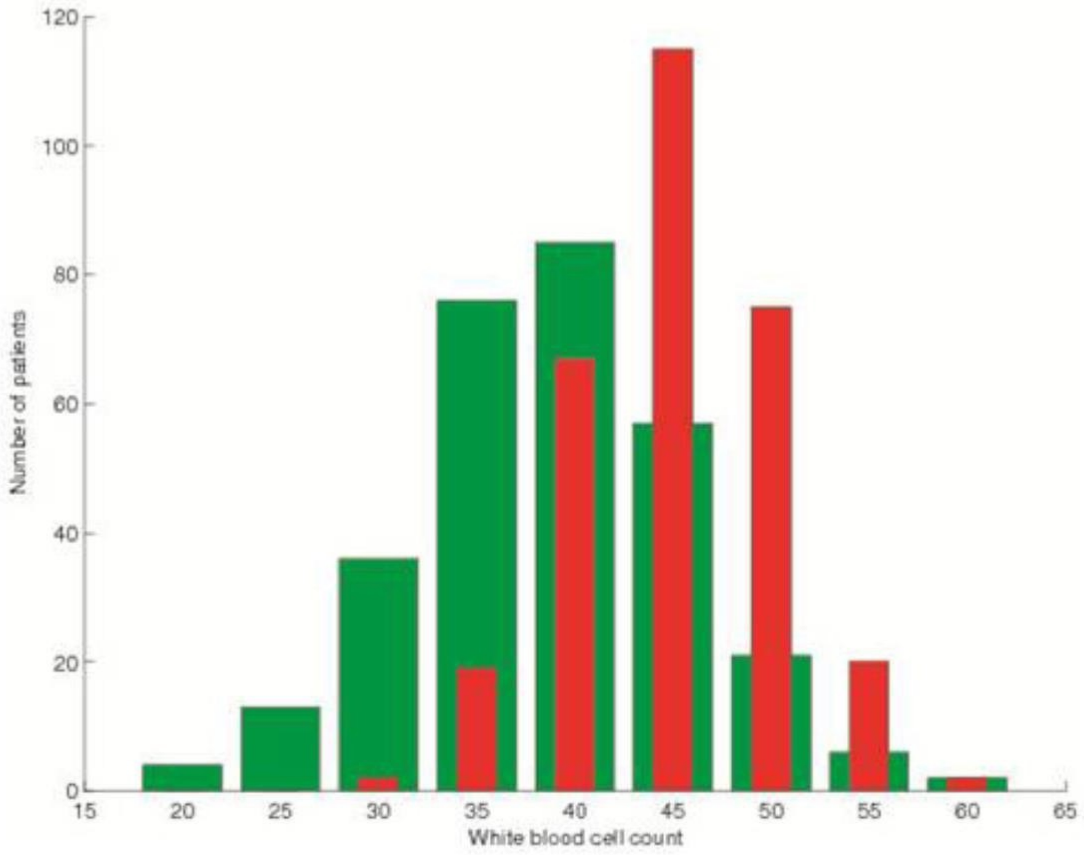
- Simple Bayes classifier: estimate posterior probability of each class

- What should the decision criterion be?
  - Choose class with highest $p(\mathcal{C}_k|\boldsymbol{x})$

- The optimal decision is the one that minimizes the expected number of mistakes

Diabetes example

- white blood cell count



This example has been adopted from Sanja Fidler's slides, University of Toronto, CSC411

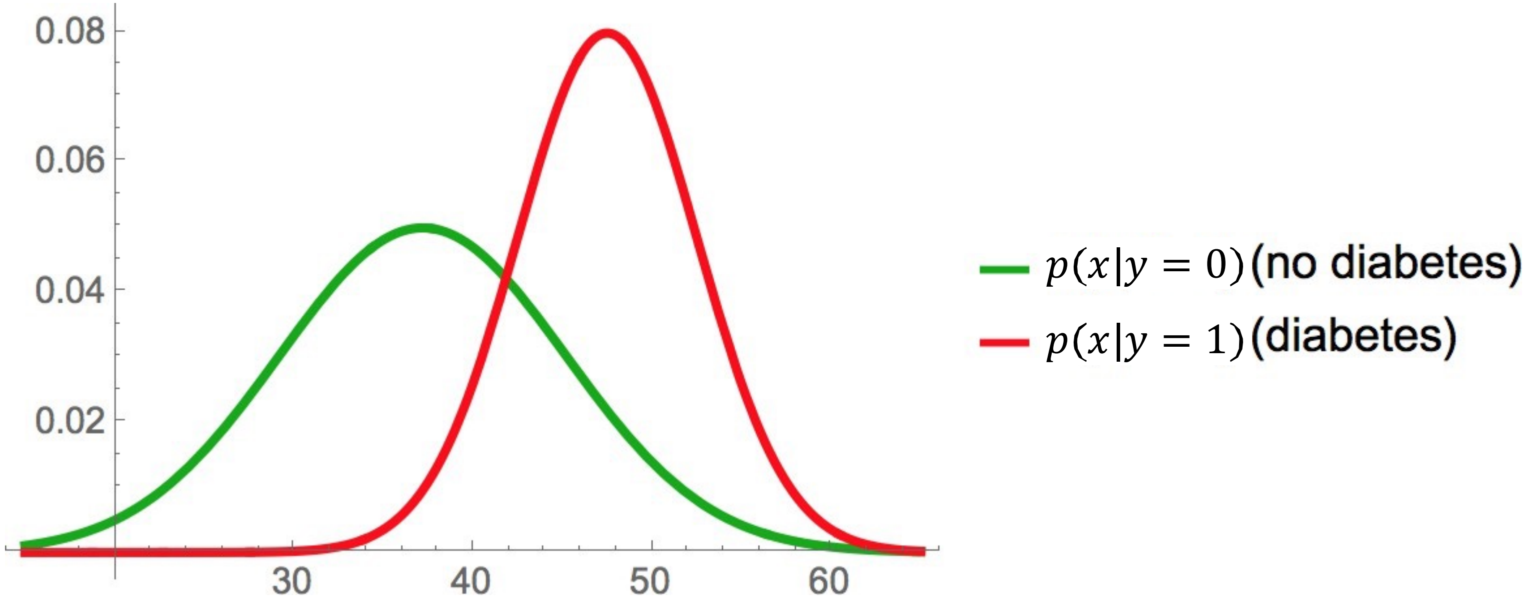
Diabetes example

- Doctor has a prior $p(y = 1) = 0.2$
  - Prior: In the absence of any observation, what do I know about the probability of the classes?

- A patient comes in with white blood cell count x

- Does the patient have diabetes $p(y = 1|x)$?
  - given a new observation, we still need to compute the posterior

Diabetes example



This example has been adopted from Sanja Fidler's slides, University of Toronto, CSC411

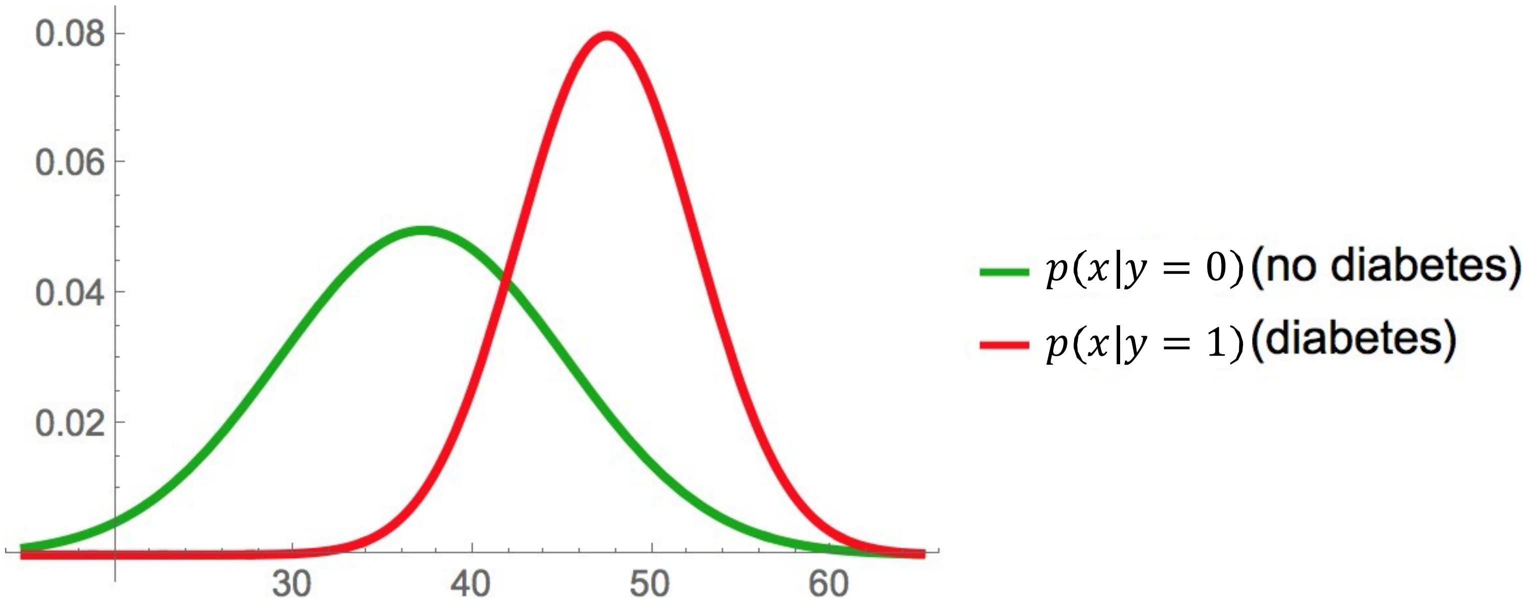
Estimate probability densities from data

- If we assume Gaussian distributions for $p(x|\mathcal{C}_1)$ and $p(x|\mathcal{C}_2)$

- Recall that for samples $\{x^{(1)}, \ldots, x^{(N)}\}$, if we assume a Gaussian distribution, the MLE estimates will be

$$
\begin{aligned}
\mu &= \frac{1}{N}\sum_{n=1}^{N} x^{(n)} \\
\sigma^2 &= \frac{1}{N}\sum_{n=1}^{N} (x^{(n)} - \mu)^2
\end{aligned}
$$

Diabetes example



$$p(x|y=1) = N(\mu_1, \sigma_1^2)$$

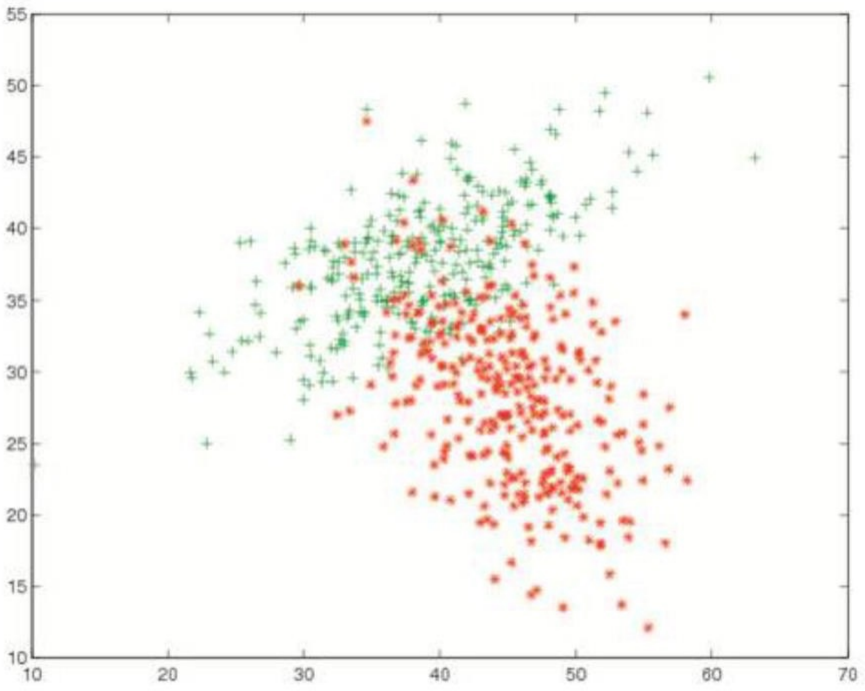
$$\mu_1 = \frac{\sum_{n:\, y^{(n)}=1} x^{(n)}}{{}_{n:\, y^{(n)}=1} 1} = \frac{\sum_{n:\, y^{(n)}=1} x^{(n)}}{N_1}$$

$$\sigma_1^2 = \frac{\sum_{n:\, y^{(n)}=1} \left(x^{(n)} - \mu_1\right)^2}{N_1}$$

This example has been adopted from Sanja Fidler's slides, University of Toronto, CSC411

Diabetes example

- Add a second observation: Plasma glucose value



Generative approach for this example

- Multivariate Gaussian distributions for $p(x|\mathcal{C}_k)$:

$$p(\boldsymbol{x}|y=k) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\}$$

$$k = 1,2$$

- Prior distribution $p(x|\mathcal{C}_k)$:
  - $p(y=1)=\pi, \qquad p(y=0)=1-\pi$

MLE for multivariate Gaussian

- For samples $\{x^{(1)}, \ldots, x^{(N)}\}$, if we assume a multivariate Gaussian distribution, the MLE estimates will be:

$$\boldsymbol{\mu} = \frac{\sum_{n=1}^{N} \boldsymbol{x}^{(n)}}{N}$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}\right)\left(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}\right)^{T}$$

Correlation matrix

$$\boldsymbol{X} = \begin{bmatrix} x_1^{(1)} & \cdots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}$$

$$\frac{1}{N}\boldsymbol{X}^T\boldsymbol{X} = \frac{1}{N}\begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(N)} \\ \vdots & \ddots & \vdots \\ x_d^{(1)} & \cdots & x_d^{(N)} \end{bmatrix}\begin{bmatrix} x_1^{(1)} & \cdots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}$$

$$= \frac{1}{N}\begin{bmatrix} \sum_{n=1}^{N} x_1^{(n)} x_1^{(n)} & \cdots & \sum_{n=1}^{N} x_1^{(n)} x_d^{(n)} \\ \vdots & \ddots & \vdots \\ \sum_{n=1}^{N} x_d^{(n)} x_1^{(n)} & \cdots & \sum_{n=1}^{N} x_d^{(n)} x_d^{(n)} \end{bmatrix}$$

Covariance Matrix

$$\boldsymbol{\mu_x} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} = \begin{bmatrix} E(x_1) \\ \vdots \\ E(x_d) \end{bmatrix}$$

$$\boldsymbol{\Sigma} = E\left[(\boldsymbol{x} - \boldsymbol{\mu_x})(\boldsymbol{x} - \boldsymbol{\mu_x})^T\right]$$

- ML estimate of covariance matrix from data points $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$:

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}\right)\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}\right)^T = \frac{1}{N}\left(\boldsymbol{X}^T\boldsymbol{X}\right)$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \vdots \\ \boldsymbol{x}^{(N)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}^{(1)} - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{x}^{(N)} - \boldsymbol{\mu} \end{bmatrix} \qquad \boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}^{(i)}$$

Mean-centered data

Generative approach: example

Maximum likelihood estimation ($D = \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$):

- $\pi = \frac{N_1}{N}$
- $\boldsymbol{\mu}_1 = \frac{\sum_{n=1}^{N} y^{(n)} \boldsymbol{x}^{(n)}}{N_1}, \boldsymbol{\mu}_2 = \frac{\sum_{n=1}^{N} (1-y^{(n)}) \boldsymbol{x}^{(n)}}{N_2}$
- $\boldsymbol{\Sigma}_1 = \frac{1}{N_1} \sum_{n=1}^{N} y^{(n)} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu})(\boldsymbol{x}^{(n)} - \boldsymbol{\mu})^T$
- $\boldsymbol{\Sigma}_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1 - y^{(n)}) (\boldsymbol{x}^{(n)} - \boldsymbol{\mu})(\boldsymbol{x}^{(n)} - \boldsymbol{\mu})^T$

$N_1 = \sum_{n=1}^{N} y^{(n)}$

$N_2 = N - N_1$

Decision boundary for Gaussian Bayes classifier

$$p(\mathcal{C}_1|\boldsymbol{x}) = p(\mathcal{C}_2|\boldsymbol{x})$$

$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\boldsymbol{x})}$$

$$\ln p(\mathcal{C}_1|\boldsymbol{x}) = \ln p(\mathcal{C}_2|\boldsymbol{x})$$

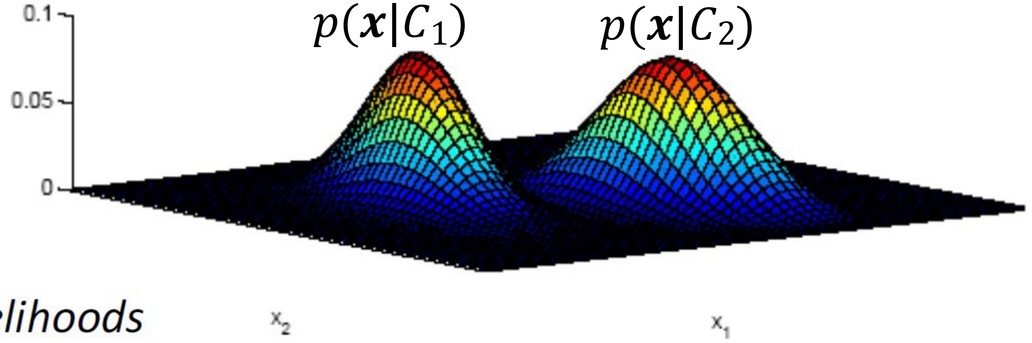
$$\ln p(\boldsymbol{x}|\mathcal{C}_1) + \ln p(\mathcal{C}_1) - \ln p(\boldsymbol{x})$$
$$= \ln p(\boldsymbol{x}|\mathcal{C}_2) + \ln p(\mathcal{C}_2) - \ln p(\boldsymbol{x})$$

$$\ln p(\boldsymbol{x}|\mathcal{C}_1) + \ln p(\mathcal{C}_1) = \ln p(\boldsymbol{x}|\mathcal{C}_2) + \ln p(\mathcal{C}_2)$$

$$\ln p(\boldsymbol{x}|\mathcal{C}_k)$$
$$= -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k^{-1}| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

Decision boundary

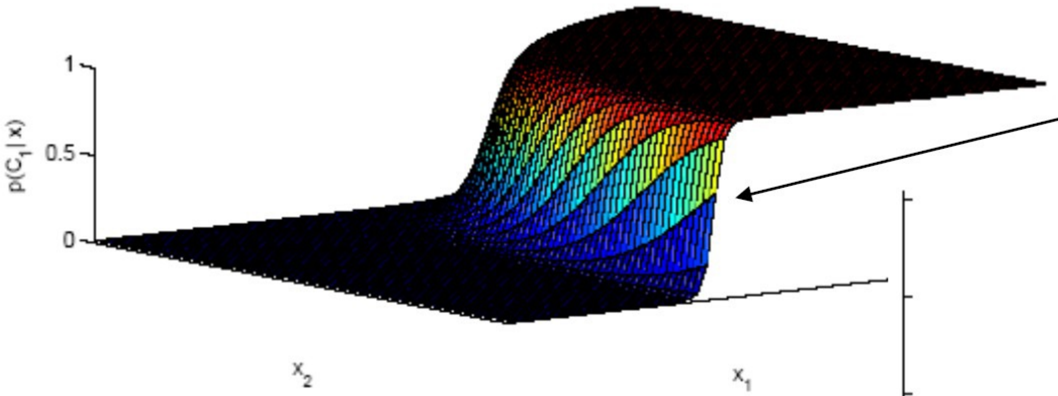
$p(\boldsymbol{x}|C_1)$ $p(\boldsymbol{x}|C_2)$



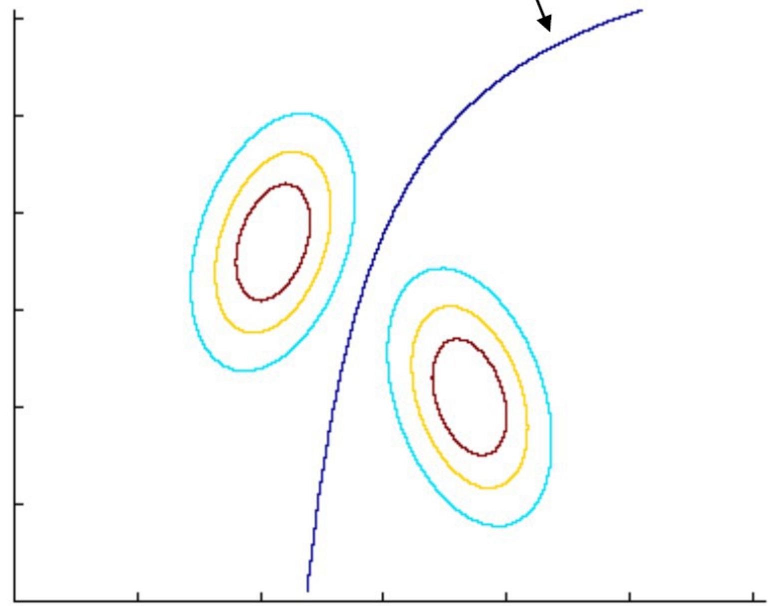
likelihoods

discriminant:

$p(C_1|\boldsymbol{x})$=$p(C_2|\boldsymbol{x})$



posterior $p(C_1|\boldsymbol{x})$



Shared covariance matrix

- When classes share a single covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$

$$p(\boldsymbol{x}|C_k) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\}$$

$$k = 1,2$$

- $p(C_1) = \pi, \qquad p(C_2) = 1 - \pi$

Likelihood

$$\prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}, y^{(n)} | \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

$$= \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)} | y^{(n)}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) p(y^{(n)} | \pi)$$

Shared covariance matrix

- Maximum likelihood estimation ($D = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$):

$$\pi = \frac{N_1}{N}$$

$$\boldsymbol{\mu}_1 = \frac{\sum_{n=1}^{N} y^{(n)}\boldsymbol{x}^{(n)}}{N_1}$$

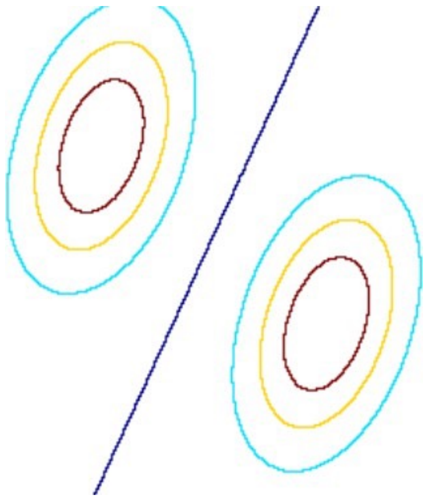
$$\boldsymbol{\mu}_2 = \frac{\sum_{n=1}^{N} (1 - y^{(n)})\boldsymbol{x}^{(n)}}{N_2}$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\left(\sum_{n \in C_1} \left(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_1\right)\left(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_1\right)^T + \sum_{n \in C_2} \left(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_2\right)\left(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_2\right)^T\right)$$

Decision boundary when shared covariance matrix

$$\ln p(\boldsymbol{x}|\mathcal{C}_1) + \ln p(\mathcal{C}_1) = \ln p(\boldsymbol{x}|\mathcal{C}_2) + \ln p(\mathcal{C}_2)$$

$$\ln p(\boldsymbol{x}|\mathcal{C}_k) = -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_k^{-1}\right| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$$



Naïve Bayes classifier

- Generative methods
  - High number of parameters

- Assumption: Conditional independence

$$p(\boldsymbol{x}|C_k) = p(x_1|C_k) \times p(x_2|C_k) \times \cdots \times p(x_d|C_k)$$

Naïve Bayes classifier

- In the decision phase, it finds the label of $\boldsymbol{x}$ according to:

$$\underset{k=1,\ldots,K}{\operatorname{argmax}}\, p(C_k|\boldsymbol{x})$$

$$\underset{k=1,\ldots,K}{\operatorname{argmax}}\, p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

$$p(\boldsymbol{x}|C_k) = p(x_1|C_k) \times p(x_2|C_k) \times \cdots \times p(x_d|C_k)$$

$$p(C_k|\boldsymbol{x}) \propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

Naïve Bayes: discrete example

- $p(H = Yes) = 0.3$

- $p(D = Yes|H = Yes) = \frac{1}{3}$
- $p(S = Yes|H = Yes) = \frac{2}{3}$

- $p(D = Yes|H = No) = \frac{2}{7}$
- $p(S = Yes|H = No) = \frac{2}{7}$

Diabetes (D)	Smoke (S)	Heart Disease (H)
Y	N	Y
Y	N	N
N	Y	N
N	Y	N
N	N	N
N	Y	Y
N	N	N
N	Y	Y
N	N	N
Y	N	N

- Decision on $\boldsymbol{x} = [Yes, Yes]$ (a person that has diabetes and also smokes):
  - $p(H = Yes|\boldsymbol{x}) \propto p(H = Yes)p(D = yes|H = Yes)p(S = yes|H = Yes) = 0.066$
  - $p(H = No|\boldsymbol{x}) \propto p(H = No)p(D = yes|H = No)p(S = yes|H = No) = 0.057$
  - Thus decide $H = yes$

Probabilistic classifiers

- How can we find the probabilities required in the Bayes decision rule?

- Probabilistic classification approaches can be divided in two main categories:

  Generative

  - Estimate pdf $p(\boldsymbol{x}, \mathcal{C}_k)$ for each class $\mathcal{C}_k$ and then use it to find $p(\mathcal{C}_k|\boldsymbol{x})$
    - or alternatively estimate both pdf $p(\boldsymbol{x}|\mathcal{C}_k)$ and $p(\mathcal{C}_k)$ to find $p(\mathcal{C}_k|\boldsymbol{x})$

  Discriminative

  - Directly estimate $p(\mathcal{C}_k|\boldsymbol{x})$ for each class $\mathcal{C}_k$

Generative approach

- <u>Inference stage</u>
  - Determine class conditional densities $p(\boldsymbol{x}|\mathcal{C}_k)$ and priors $p(\mathcal{C}_k)$
  - Use the Bayes theorem to find $p(\mathcal{C}_k|\boldsymbol{x})$

- <u>Decision stage</u>: After learning the model (inference stage), make optimal class assignment for new input
  - if $p(\mathcal{C}_i|\boldsymbol{x}) > p\left(\mathcal{C}_j\middle|\boldsymbol{x}\right) \quad \forall j \neq i$ then decide $\mathcal{C}_i$

Resource

- Yaser S. Abu-Mostafa, MalikMaghdon-Ismail, and Hsuan Tien Lin,“**Learning from Data**”, 2012.
- C. Bishop, “Pattern Recognition and Machine Learning”, Chapter 2.