Probabilistic classification

Machine Learning

Hamid R Rabiee – Zahra Dehghanian

Spring 2025

Sharif University
of Technology

Topics

- Probabilistic approach
  - Bayes decision theory
  - Generative models
    - Gaussian Bayes classifier
    - Naïve Bayes

Classification problem: probabilistic view

- Given: Training set
  - labeled set of N input-output pairs $D = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$
  - $y \in \{1, \dots, K\}$
- Goal: Given an input $\boldsymbol{x}$, assign it to one of K classes
- Examples:
  - Spam filter
  - Handwritten digit recognition
  - …

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 $\boldsymbol{x}$ $(p(\boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k))$

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

$K = 2$

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

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

- 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 decision rule
Minimizing misclassification rate

$K = 2$

- Decision regions: $\mathcal{R}_k = \{\boldsymbol{x} | \alpha(\boldsymbol{x}) = k\}$
  - All points in $\mathcal{R}_k$ are assigned to class $\mathcal{C}_k$

$$p(error) = E_{\boldsymbol{x},y}[I(\alpha(\boldsymbol{x}) \neq y)]$$

$$= p(\boldsymbol{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\boldsymbol{x} \in \mathcal{R}_2, \mathcal{C}_1)$$

$$= \int_{\mathcal{R}_1} p(\boldsymbol{x}, \mathcal{C}_2)\, d\boldsymbol{x} + \int_{\mathcal{R}_2} p(\boldsymbol{x}, \mathcal{C}_1)\, d\boldsymbol{x}$$

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

Choose class with highest $p(\mathcal{C}_k|\boldsymbol{x})$ as $\alpha(\boldsymbol{x})$

Bayes minimum error

- Bayes minimum error classifier:

$$\min_{\alpha(.)} E_{\boldsymbol{x},y}[I(\alpha(\boldsymbol{x}) \neq y)]$$

Zero-one loss

  - If we know the probabilities in advance then the above optimization problem will be solved easily.
    - $\alpha(\boldsymbol{x}) = \underset{y}{\operatorname{argmax}}\, p(y|\boldsymbol{x})$

- In practice, we can estimate $p(y|\boldsymbol{x})$ based on a set of training samples $\mathcal{D}$

Bayes theorem

- Bayes' theorem

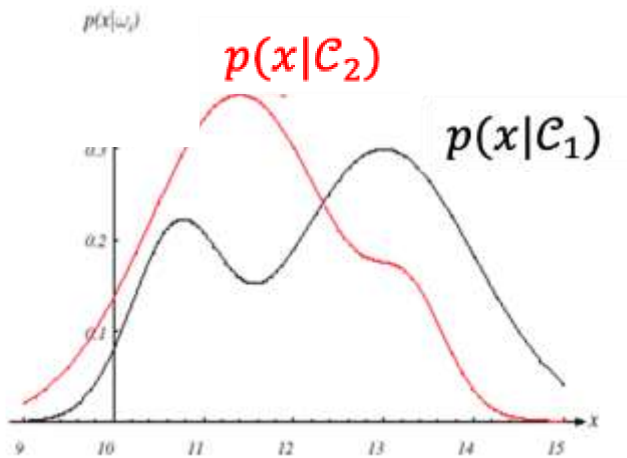
$$\underbrace{p(\mathcal{C}_k|\boldsymbol{x})}_{\text{Posterior}} = \frac{\overbrace{p(\boldsymbol{x}|\mathcal{C}_k)}^{\text{Likelihood}}\overbrace{p(\mathcal{C}_k)}^{\text{Prior}}}{p(\boldsymbol{x})}$$

- 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 $\mathcal{C}_k$

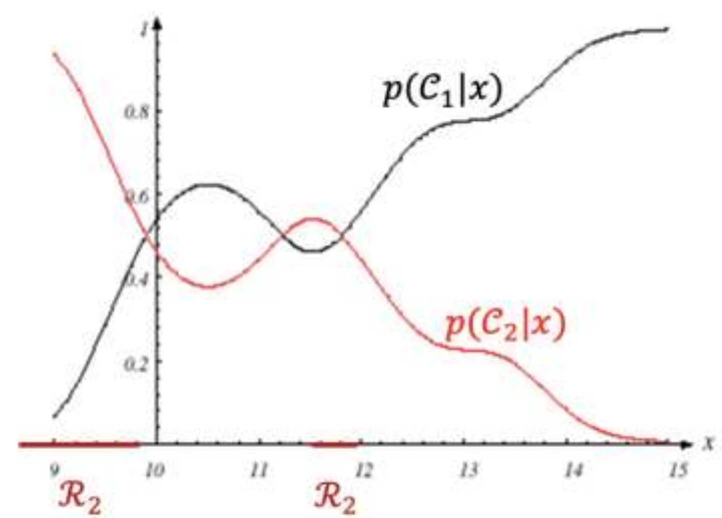
Bayes decision rule: example

- 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 $P(C_1|\boldsymbol{x}) > P(C_2|\boldsymbol{x})$ decide C_1

  otherwise decide C_2

Equivalent

- If $\frac{p(\boldsymbol{x}|C_1)P(C_1)}{p(\boldsymbol{x})} > \frac{p(\boldsymbol{x}|C_2)P(C_2)}{p(\boldsymbol{x})}$ decide C_1

  otherwise decide C_2

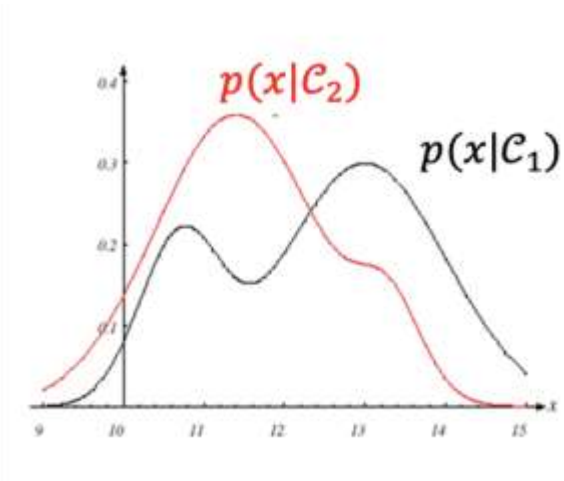
Equivalent

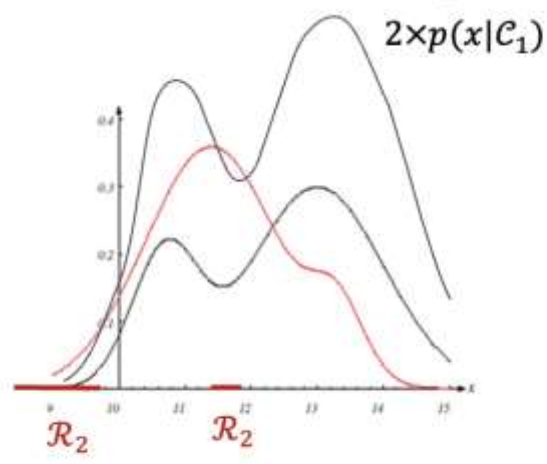
- If $p(\boldsymbol{x}|C_1)P(C_1) > p(\boldsymbol{x}|C_2)P(C_2)$ decide C_1

  otherwise decide C_2

Bayes decision rule: example

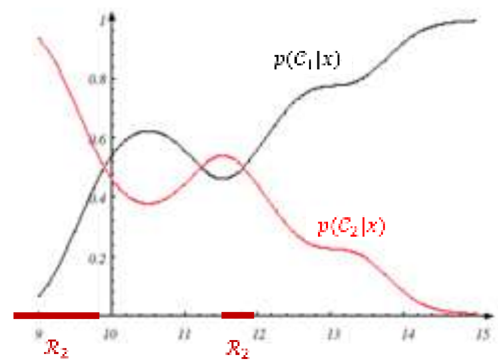
- Bayes decision: Choose the class with highest $p(\mathcal{C}_k|\boldsymbol{x})$





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

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



Bayes Classier

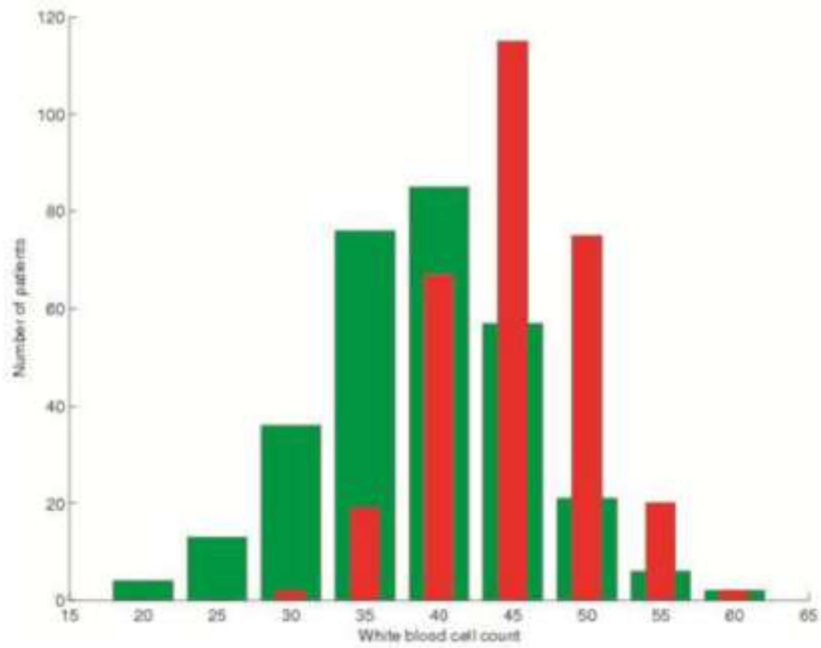
- Simple Bayes classifier: estimate posterior probability of each class

- 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

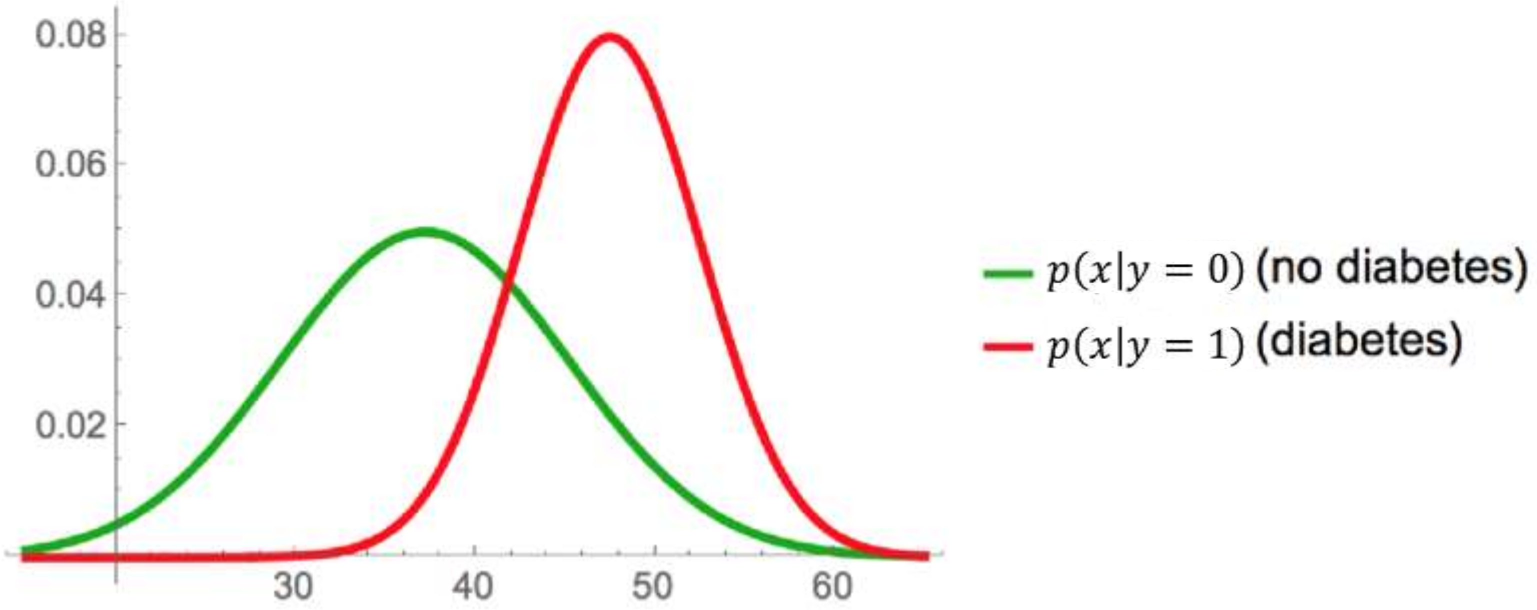
- Doctor has a prior $p(y = 1) = 0.2$
  - Prior: In the absence of any observation, what do I know about the probability of the classes?

- 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

$$p(x = 40|y = 0)P(y = 0) >^{?} p(x = 40|y = 1)P(y = 1)$$



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

Estimate probability densities from data

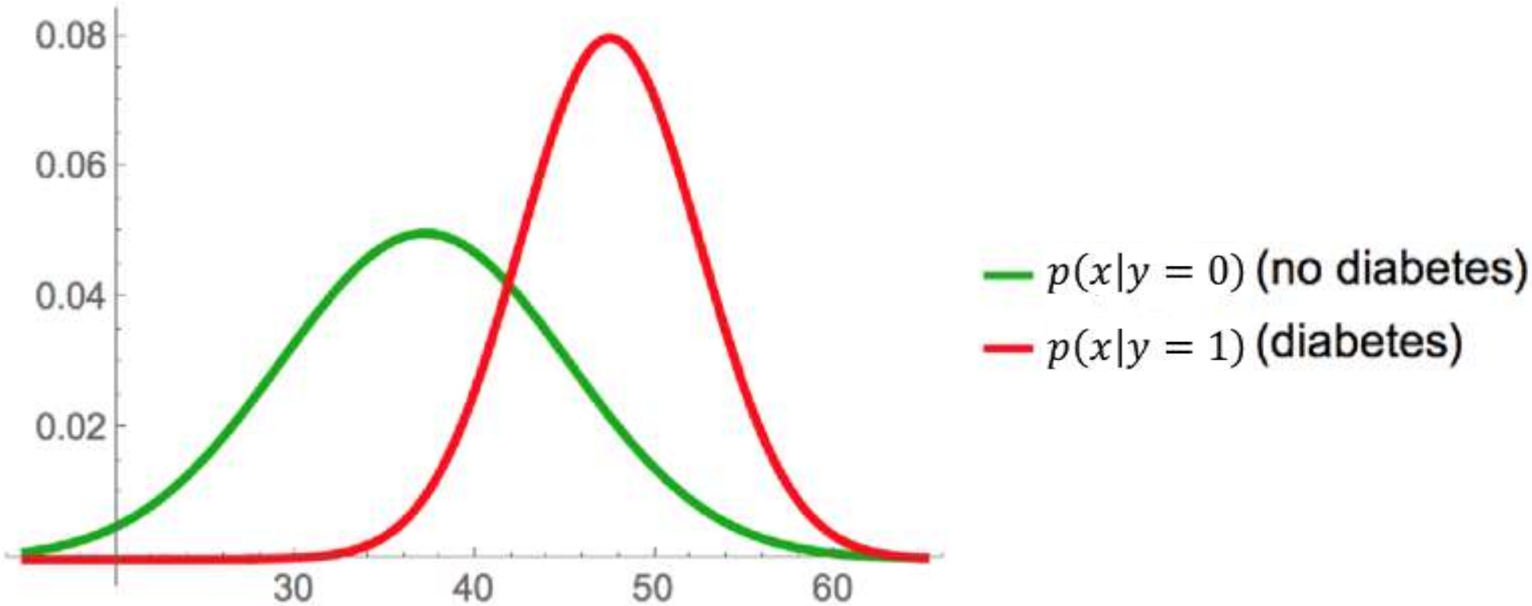
- If we assume Gaussian distributions for $p(x|y=0)$ and $p(x|y=1)$

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

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

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

Diabetes example



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

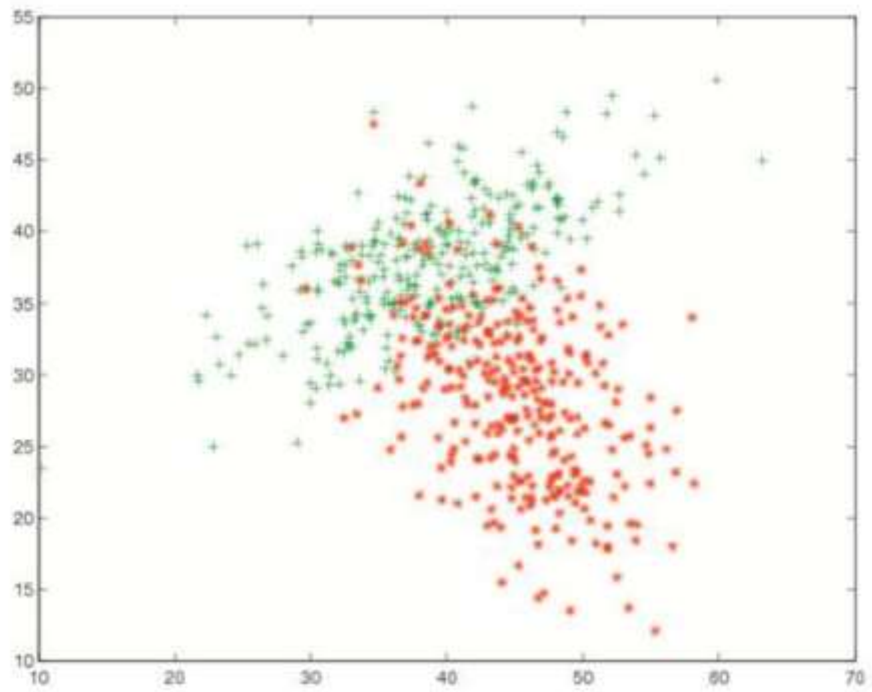
$$\mu_1 = \frac{\sum_{n:\, y^{(n)}=1} x^{(n)}}{\sum_{n:\, y^{(n)}=1} 1} = \frac{\sum_{n:\, y^{(n)}=1} x^{(n)}}{N_1}$$

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

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

Diabetes example

- Add a second observation: Plasma glucose value



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

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:

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

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

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

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

Naïve Bayes: discrete example

- $p(h) = 0.3$

- $p(d|h) = \frac{1}{3}$
- $p(s|h) = \frac{2}{3}$

- $p(d|\bar{h}) = \frac{2}{7}$
- $p(s|\bar{h}) = \frac{2}{7}$

$H = Yes \quad \equiv h$
$H = No \quad \equiv \bar{h}$

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

Naïve Bayes: discrete example

- $p(h) = 0.3$

- $p(d|h) = \frac{1}{3}$
- $p(s|h) = \frac{2}{3}$

- $p(d|\bar{h}) = \frac{2}{7}$
- $p(s|\bar{h}) = \frac{2}{7}$

$H = Yes \quad \equiv h$

$H = No \quad \equiv \bar{h}$

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 $\boldsymbol{x} = [d, \bar{s}]$ (a person that has diabetes but does not smoke):
  - $p(h|\boldsymbol{x}) \propto p(h)p(d|h)p(\bar{s}|h) = 1/14$
  - $p(\bar{h}|\boldsymbol{x}) \propto p(\bar{h})p(d|\bar{h})p(\bar{s}|\bar{h}) = 1/6$
  - Thus decide $H = No$

Naïve Bayes classifier

- Finds d univariate distributions $p(x_1|C_k), \cdots, p(x_d|C_k)$ instead of finding one multi-variate distribution $p(\boldsymbol{x}|C_k)$
  - Example 1: For Gaussian class-conditional density $p(\boldsymbol{x}|C_k)$, it finds $d + d$ (mean and sigma parameters on different dimensions) instead of $d + \frac{d(d+1)}{2}$ parameters
  - Example 2: For Bernoulli class-conditional density $p(\boldsymbol{x}|C_k)$, it finds d (mean parameters on different dimensions) instead of $2^{\mathrm{d}} - 1$ parameters

- It first estimates the class conditional densities $p(x_1|C_k), \cdots, p(x_d|C_k)$ and the prior probability $p(C_k)$ for each class $(k = 1, \ldots, K)$ based on the training set.

Multivariate Gaussian

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

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

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

Multivariate Gaussian

- Multivariate Gaussian distributions for $p(x|C_k)$:

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

$$k = 1,2$$

- Prior distribution $p(y)$:
  - $p(y=1)=\pi, \qquad p(y=0)=1-\pi$

Multivariate Gaussian

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

$y \in \{0,1\}$

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

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

$$N_2 = N - N_1$$

Decision boundary for Gaussian Bayes classifier

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

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

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

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

Decision boundary for Gaussian Bayes classifier

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

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

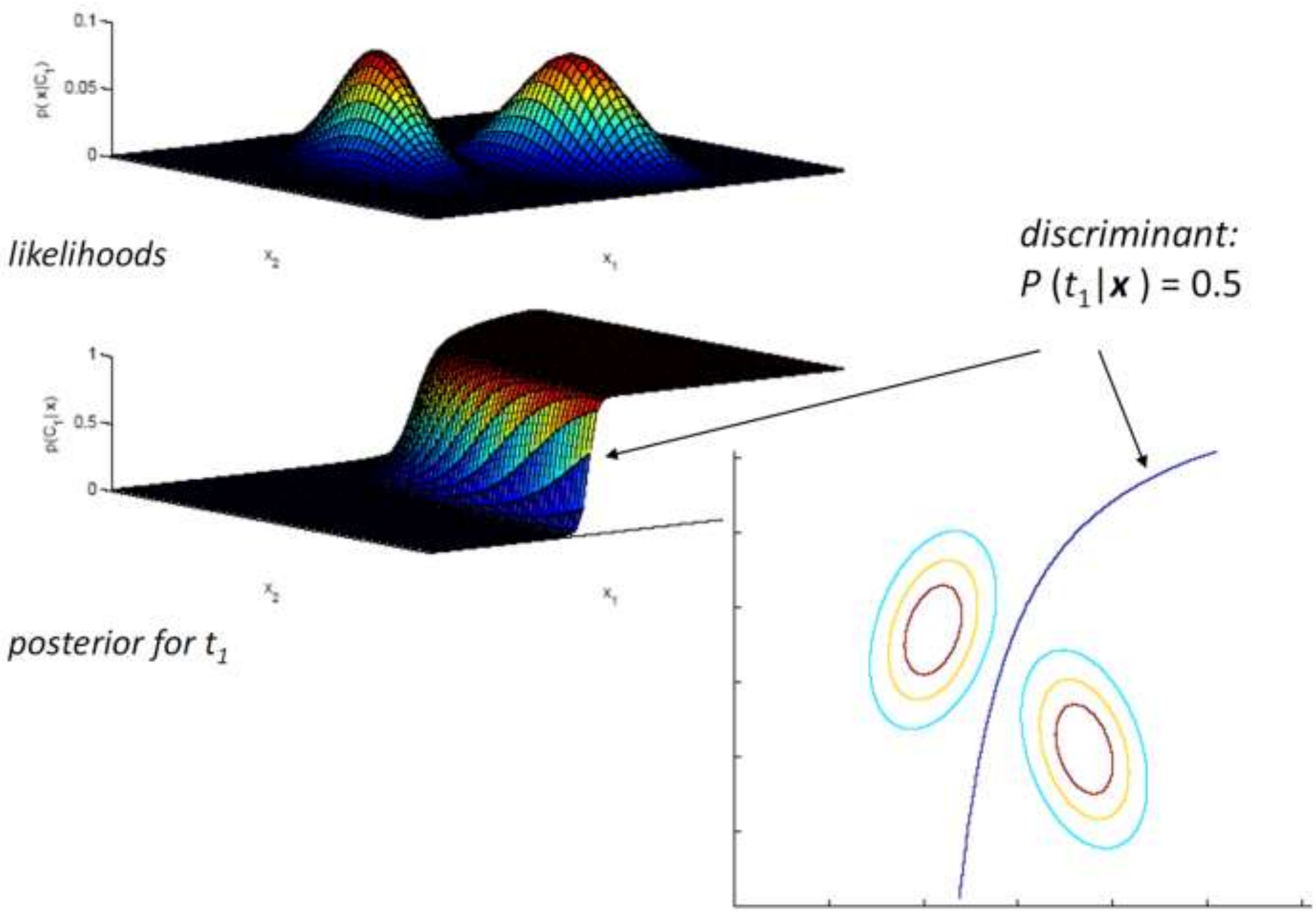
$$\ln p(\mathcal{C}_1|\boldsymbol{x}) = \ln p(\mathcal{C}_2|\boldsymbol{x})$$

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

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

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

Decision boundary



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 ($D = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$):

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

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

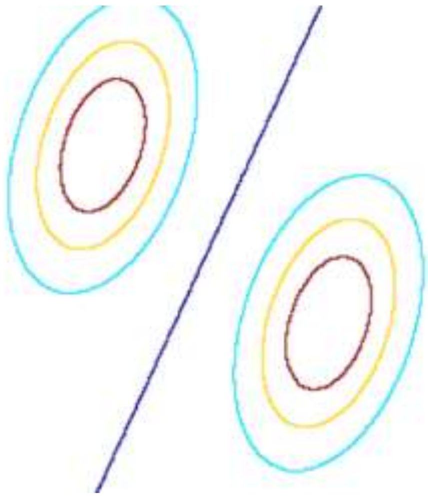
$$\boldsymbol{\mu}_2 = \frac{\sum_{n=1}^{N} (1 - y^{(n)}) \boldsymbol{x}^{(n)}}{N_2}$$

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

Decision boundary when shared covariance matrix

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

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



Multi-class Bayes decision rule

- Multi-class problem: Probability of error of Bayesian decision rule
  - Simpler to compute the probability of correct decision

$$P(error) = 1 - P(correct)$$

$$P(Correct) = \sum_{i=1}^{K} \int_{\mathcal{R}_i} p(\boldsymbol{x}, \mathcal{C}_i)\, d\boldsymbol{x}$$

$$= \sum_{i=1}^{K} \int_{\mathcal{R}_i} p(\mathcal{C}_i|\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x}$$

$\mathcal{R}_i$: the subset of feature space assigned to the class $\mathcal{C}_i$ using the classifier

Bayes minimum error

- Bayes minimum error classifier:

$$\min_{\alpha(.)} E_{\boldsymbol{x},y}[I(\alpha(\boldsymbol{x}) \neq y)]$$

Zero-one loss

$$\alpha(\boldsymbol{x}) = \underset{y}{\operatorname{argmax}}\, p(y|\boldsymbol{x})$$

Minimizing Bayes risk (expected loss)

$$E_{\boldsymbol{x},y}[L(\alpha(\boldsymbol{x}),y)]$$
$$=\int\sum_{j=1}^{K}L(\alpha(\boldsymbol{x}),\mathcal{C}_j)p(\boldsymbol{x},\mathcal{C}_j)d\boldsymbol{x}$$

Minimizing Bayes risk (expected loss)

$$E_{\boldsymbol{x},y}[L(\alpha(\boldsymbol{x}),y)]$$

$$= \int \sum_{j=1}^{K} L(\alpha(\boldsymbol{x}), \mathcal{C}_j) p(\boldsymbol{x}, \mathcal{C}_j) d\boldsymbol{x}$$

$$= \int p(\boldsymbol{x}) \sum_{j=1}^{K} L(\alpha(\boldsymbol{x}), \mathcal{C}_j) p(\mathcal{C}_j|\boldsymbol{x}) d\boldsymbol{x}$$

for each $\boldsymbol{x}$ minimize it that is called conditional risk

Minimizing Bayes risk (expected loss)

$$E_{\boldsymbol{x},y}[L(\alpha(\boldsymbol{x}),y)]$$

$$= \int \sum_{j=1}^{K} L(\alpha(\boldsymbol{x}),\mathcal{C}_j)p(\boldsymbol{x},\mathcal{C}_j)d\boldsymbol{x}$$

$$= \int p(\boldsymbol{x}) \underbrace{\sum_{j=1}^{K} L(\alpha(\boldsymbol{x}),\mathcal{C}_j)p(\mathcal{C}_j|\boldsymbol{x})d\boldsymbol{x}}_{\text{for each } \boldsymbol{x} \text{ minimize it that is called conditional risk}}$$

- Bayes minimum loss (risk) decision rule: $\hat{\alpha}(\boldsymbol{x})$

$$\hat{\alpha}(\boldsymbol{x}) = \underset{i=1,\dots,K}{\operatorname{argmin}} \sum_{j=1}^{K} L_{ij}\, p(\mathcal{C}_j|\boldsymbol{x})$$

The loss of assigning a sample to $\mathcal{C}_i$ where the correct class is $\mathcal{C}_j$

Minimizing expected loss: special case (loss = misclassification rate)

- Problem definition for this special case:
  - If action $\alpha(\boldsymbol{x}) = i$ is taken and the true category is $\mathcal{C}_j$, then the decision is correct if $i = j$ and otherwise it is incorrect.
    - Zero-one loss function:

$$L_{ij} = 1 - \delta_{ij} = \begin{cases} 0 & i = j \\ 1 & o.w. \end{cases}$$

$$\hat{\alpha}(\boldsymbol{x}) = \underset{i=1,\ldots,K}{\operatorname{argmin}} \sum_{j=1}^{K} L_{ij}\, p(\mathcal{C}_j|\boldsymbol{x})$$

Minimizing expected loss: special case (loss = misclassification rate)

- Problem definition for this special case:
  - If action $\alpha(\boldsymbol{x}) = i$ is taken and the true category is $\mathcal{C}_j$, then the decision is correct if $i = j$ and otherwise it is incorrect.
    - Zero-one loss function:

$$L_{ij} = 1 - \delta_{ij} = \begin{cases} 0 & i = j \\ 1 & o.w. \end{cases}$$

$$\hat{\alpha}(\boldsymbol{x}) = \underset{i=1,\dots,K}{\operatorname{argmin}} \sum_{j=1}^{K} L_{ij} p(\mathcal{C}_j|\boldsymbol{x})$$

$$= \underset{i=1,\dots,K}{\operatorname{argmin}}\, 0 \times p(\mathcal{C}_i|\boldsymbol{x}) + \sum_{j \neq i} p(\mathcal{C}_j|\boldsymbol{x})$$

$$= \underset{i=1,\dots,K}{\operatorname{argmin}}\, 1 - p(\mathcal{C}_i|\boldsymbol{x}) = \underset{i=1,\dots,K}{\operatorname{argmax}}\, p(\mathcal{C}_i|\boldsymbol{x})$$

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 $p(\mathcal{C}_i|\boldsymbol{x}) > p(\mathcal{C}_j|\boldsymbol{x}) \quad \forall j \neq i$ then decide $\mathcal{C}_i$

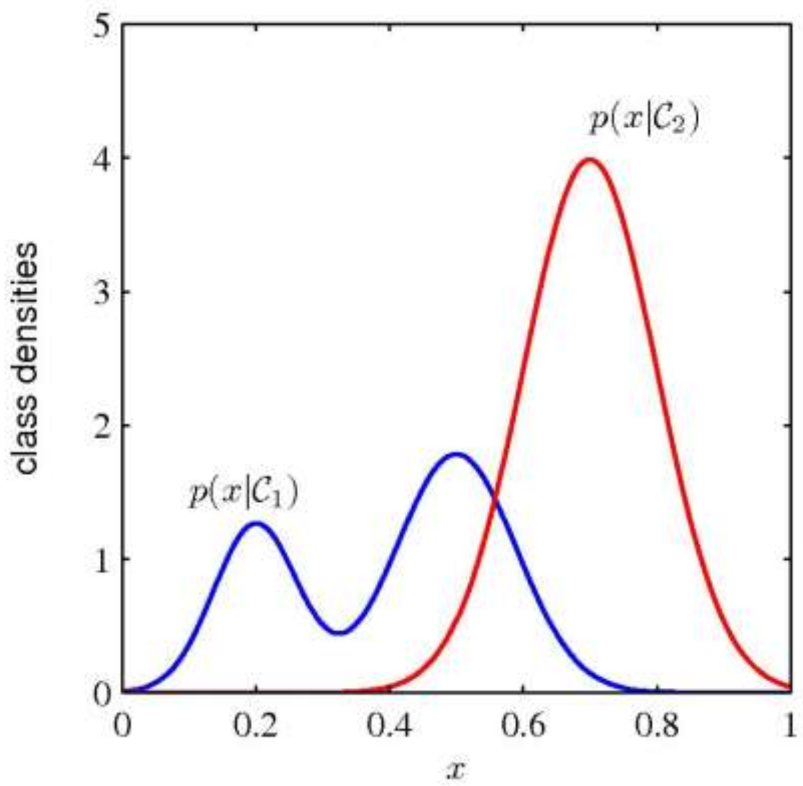
Probabilistic discriminant functions

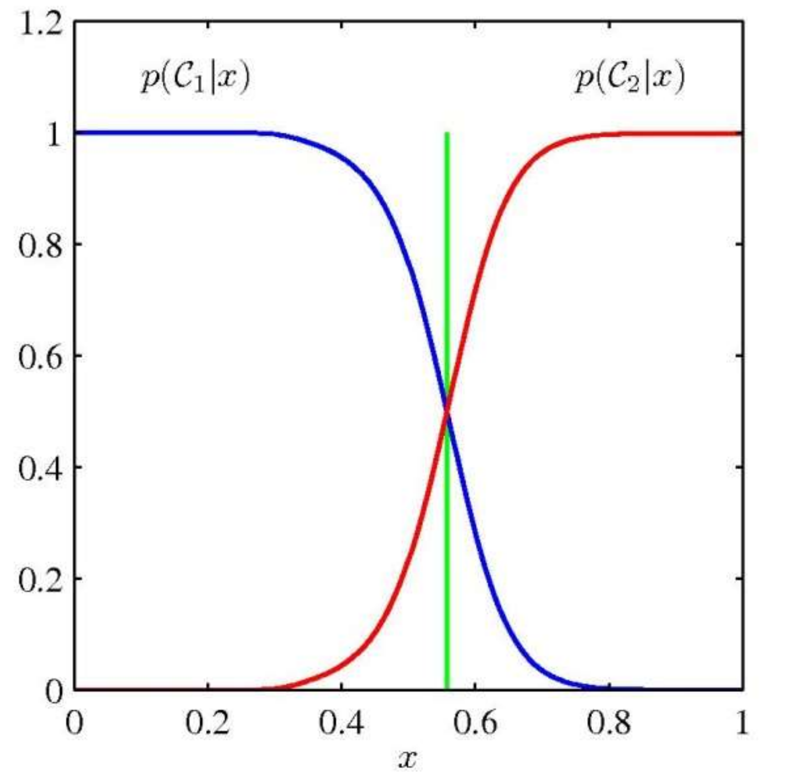
- **Discriminant functions:** A popular way of representing a classifier
  - A discriminant function $f_i(\boldsymbol{x})$ for each class $\mathcal{C}_i$ $(i = 1, \dots, K)$:
    - $\boldsymbol{x}$ is assigned to class $\mathcal{C}_i$ if:

$$f_i(\boldsymbol{x}) > f_j(\boldsymbol{x}) \quad \forall j \neq i$$

- Representing Bayesian classifier using discriminant functions:
  - Classifier minimizing error rate: $f_i(\boldsymbol{x}) = P(\mathcal{C}_i|\boldsymbol{x})$
  - Classifier minimizing risk: $f_i(\boldsymbol{x}) = -\sum_{j=1}^{K} L_{ij} p(\mathcal{C}_j|\boldsymbol{x})$

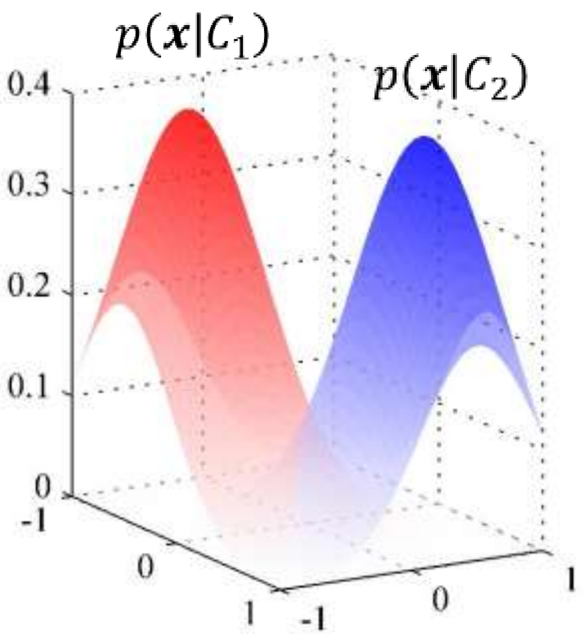
Discriminative vs. generative approach

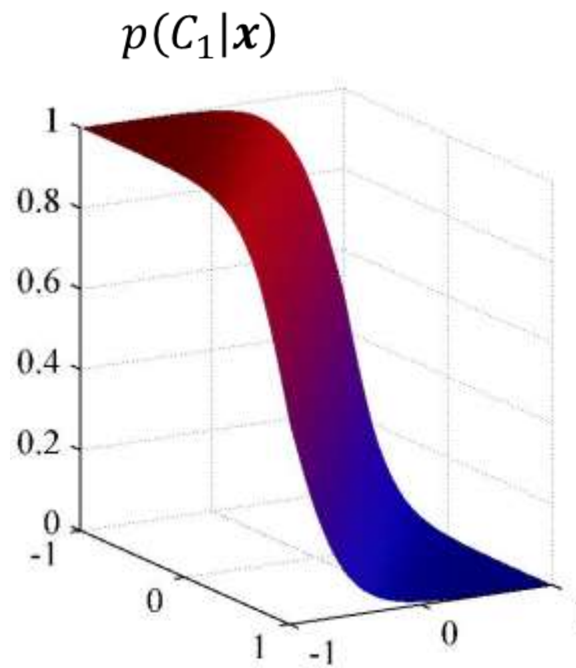




[Bishop]

Class conditional densities vs. posterior





[Bishop]

$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x} + w_0)$$

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

$$\sigma(z) = \frac{1}{1 + \exp(z)}$$

Feed back

? https://forms.gle/vKRbyVVsWRKcZuqr8

Resources

- C. Bishop, “Pattern Recognition and Machine Learning”, Chapter 4.2-4.3.
- Course CE-717, Dr. M.Soleymani