Probabilistic classifiers

CE-477: Machine Learning - CS-828: Theory of Machine Learning
Sharif University of Technology
Fall 2024

Fatemeh Seyyedsalehi

References of the lecture are mentioned in the last slide

Topics

- Probabilistic approach
  - Bayes decision theory
  - Generative models
    - Gaussian Bayes classifier
    - Naïve Bayes
  - Discriminative models
    - Logistic regression

Classification problem: probabilistic view

- Each feature as a random variable
- Class label also as a random variable
- We observe the feature values for a random sample and we intend to find its class label
  - Evidence: feature vector $\boldsymbol{x}$
  - Query: class label

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

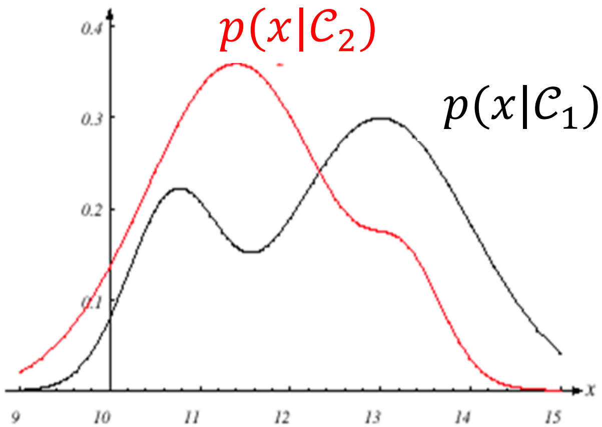
$$\underbrace{p(\mathcal{C}_k|\boldsymbol{x})}_{\text{Posterior}} = \frac{\overbrace{p(\boldsymbol{x}|\mathcal{C}_k)}^{\text{Likelihood}}\overbrace{p(\mathcal{C}_k)}^{\text{Prior}}}{p(\boldsymbol{x})}$$

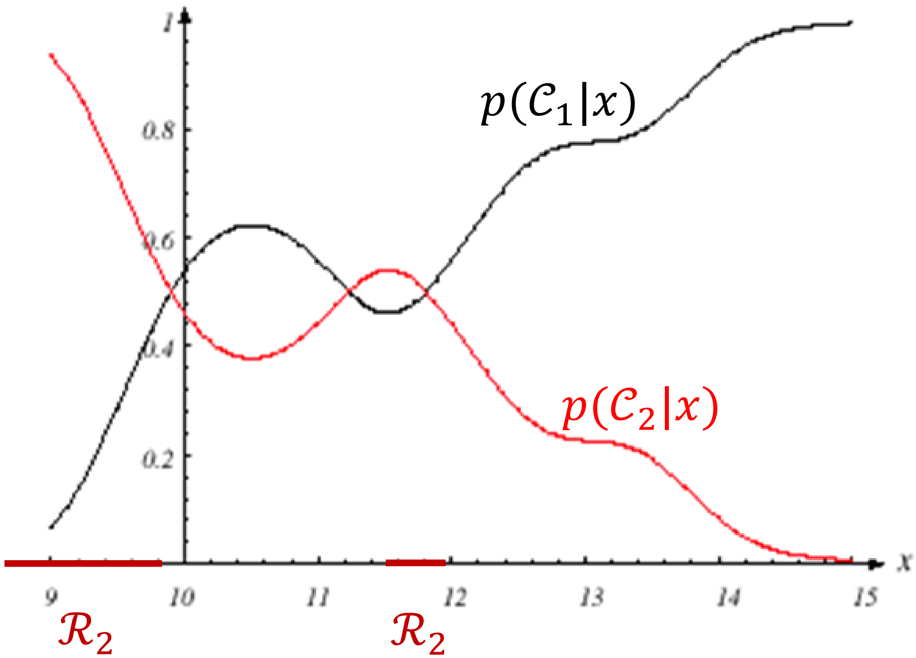
- Posterior probability: $p(\mathcal{C}_k|\boldsymbol{x})$
- Likelihood or class conditional probability: $p(\boldsymbol{x}|\mathcal{C}_k)$
- Prior probability: $p(\mathcal{C}_k)$

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

Equivalent

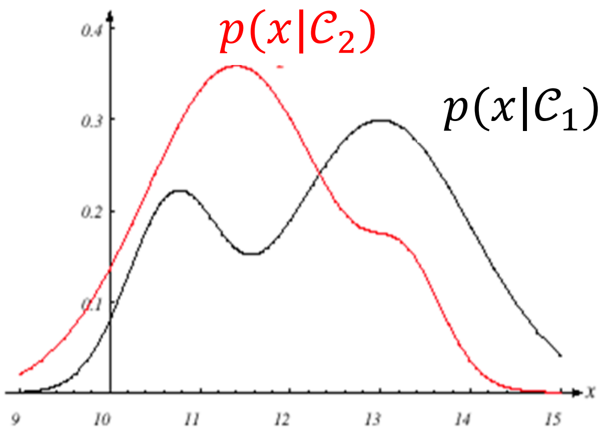
- If $\frac{p(\boldsymbol{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{p(\boldsymbol{x})} > \frac{p(\boldsymbol{x}|\mathcal{C}_2)P(\mathcal{C}_2)}{p(\boldsymbol{x})}$ decide $\mathcal{C}_1$
  otherwise decide $\mathcal{C}_2$

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 $\mathcal{C}_2$

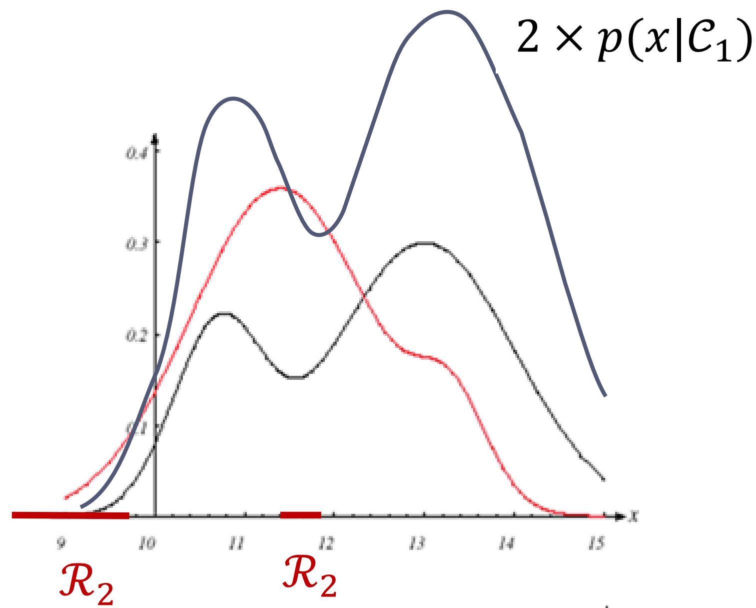
Bayes decision rule: example

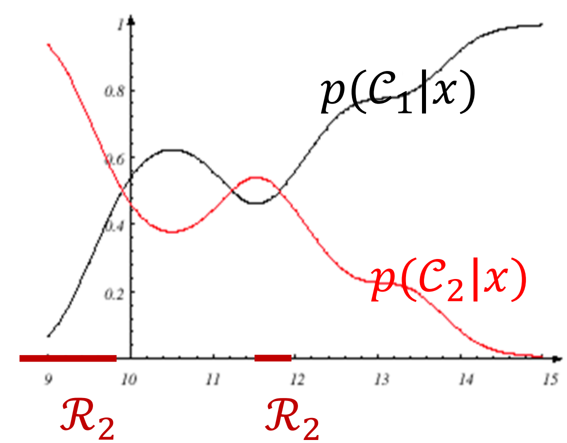
- Bayes decision: Choose the class with highest $p(\mathcal{C}_k|\boldsymbol{x})$



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

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





Bayes Classier

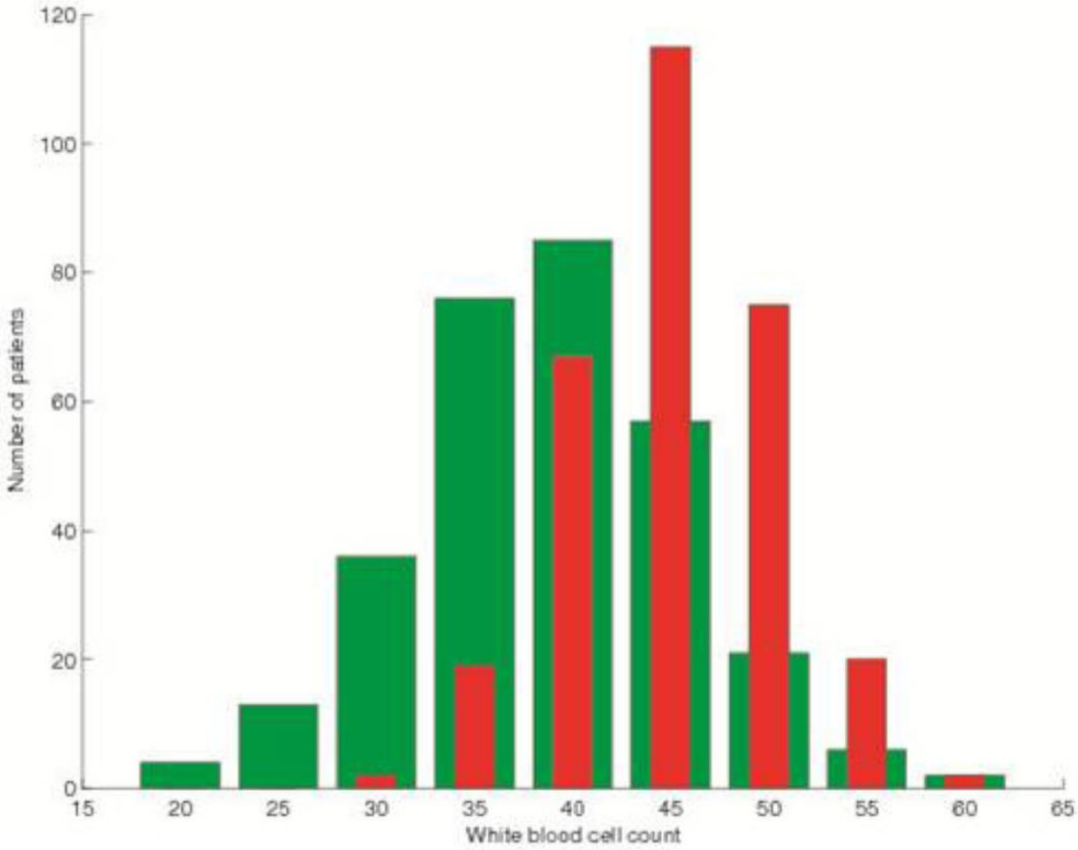
- Simple Bayes classifier: estimate posterior probability of each class

- 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

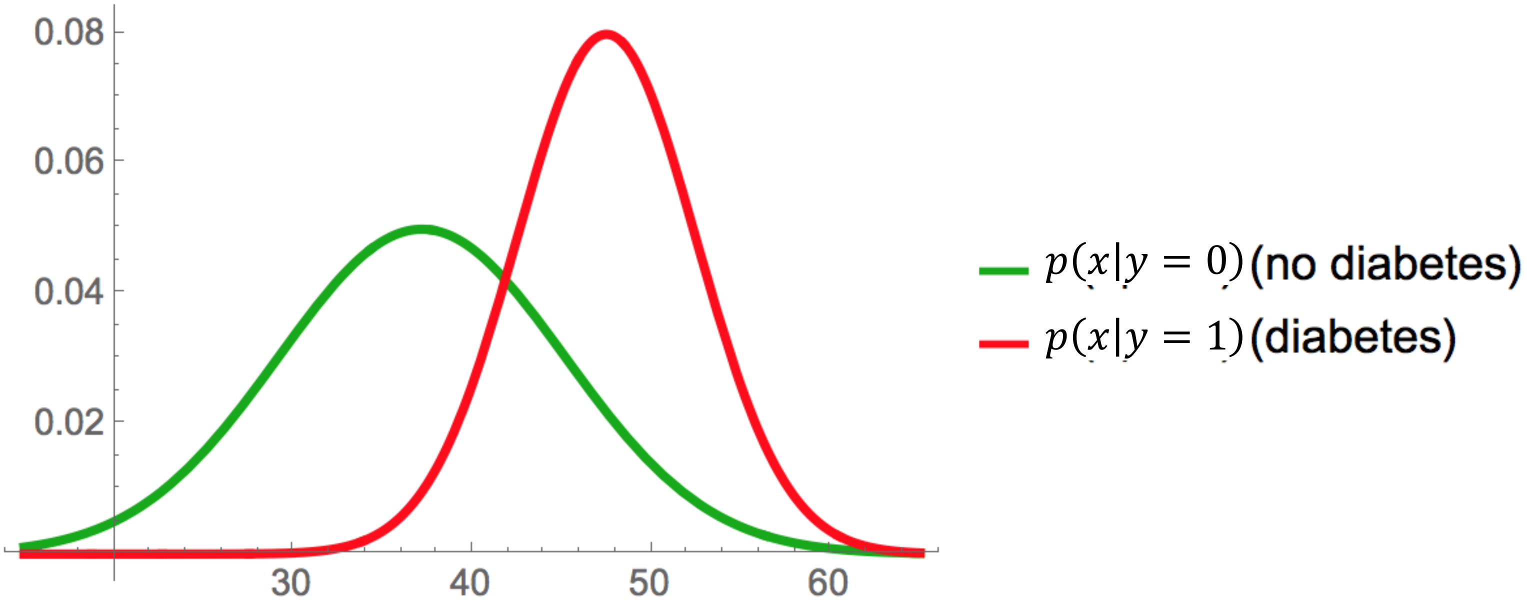
- Doctor has a prior $p(y = 1) = 0.2$
  - Prior: In the absence of any observation, what do I know about the probability of the classes?

- 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

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



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

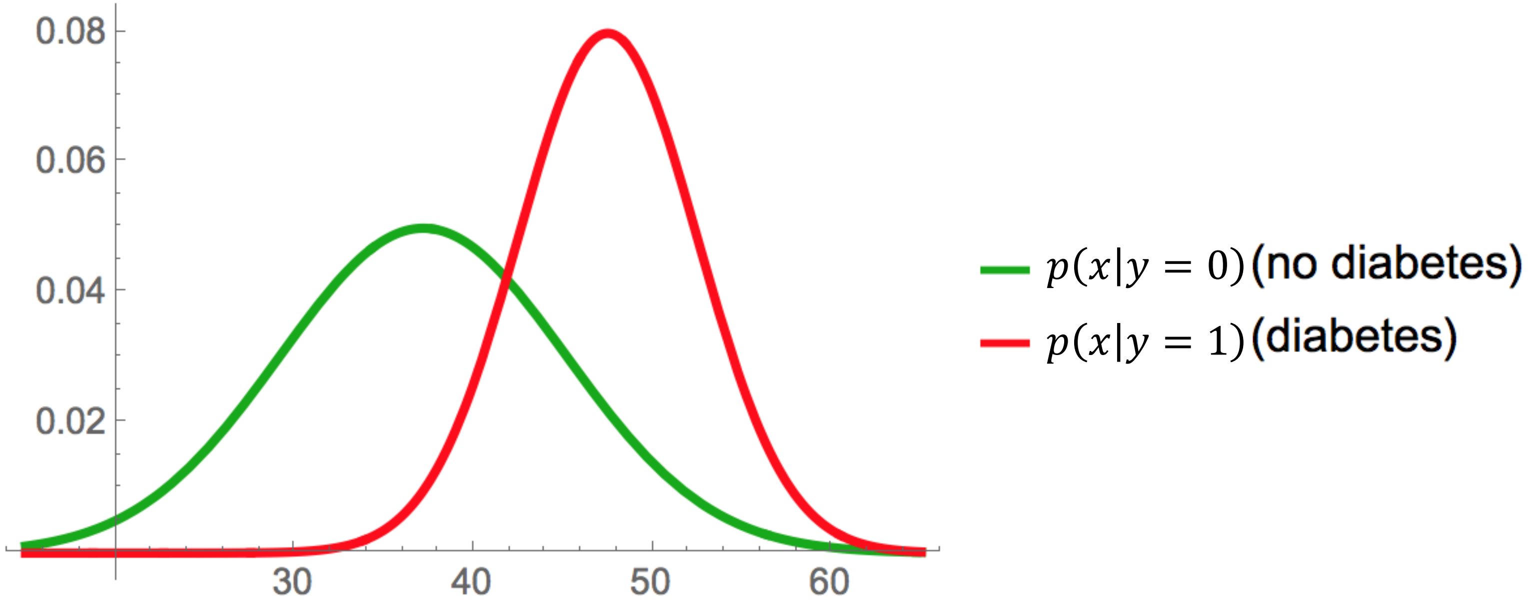
Estimate probability densities from data

- 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

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

Diabetes example



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

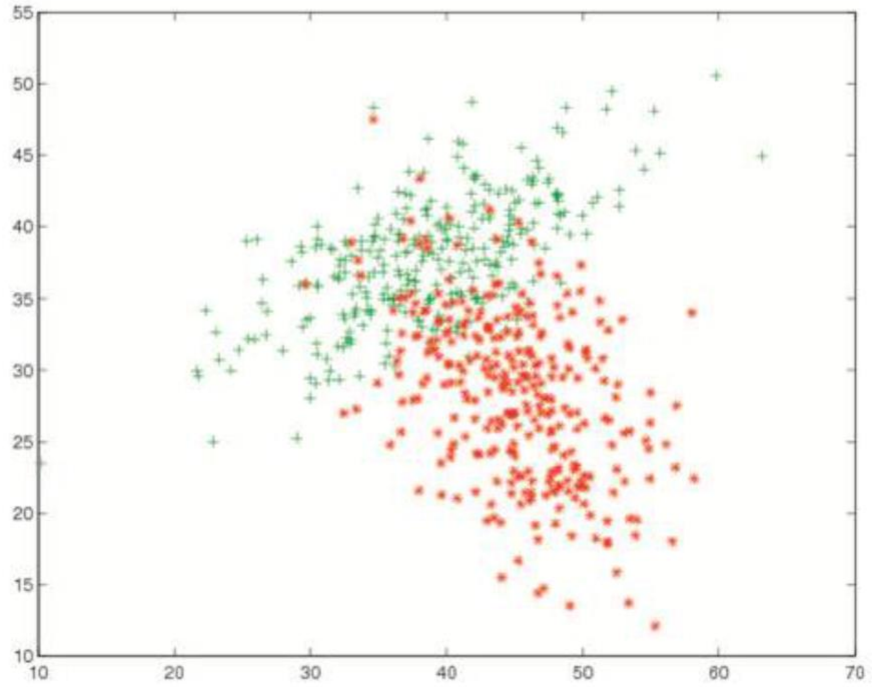
$$\mu_1 = \frac{\sum_{n:\, y^{(n)}=1} x^{(n)}}{\sum_{n:\, y^{(n)}=1} 1} = \frac{\sum_{n:\, y^{(n)}=1} x^{(n)}}{N_1}$$

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

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

Generative approach for this example

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

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

$$k = 1,2$$

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

MLE for multivariate Gaussian

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

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

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

Generative approach: example

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

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

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

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

$$N_2 = N - N_1$$

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

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

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

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

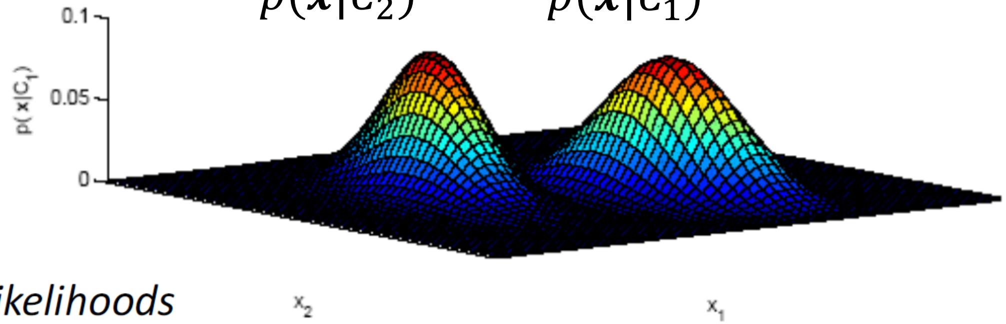
$$\begin{aligned}&\ln p(\boldsymbol{x}|\mathcal{C}_1) + \ln p(\mathcal{C}_1) - \ln p(\boldsymbol{x})\\ &= \ln p(\boldsymbol{x}|\mathcal{C}_2) + \ln p(\mathcal{C}_2) - \ln p(\boldsymbol{x})\end{aligned}$$

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

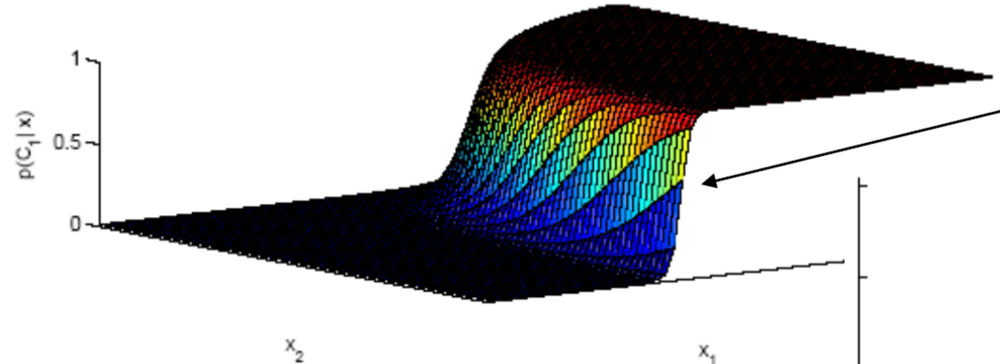
$$\begin{aligned}&\ln p(\boldsymbol{x}|\mathcal{C}_k)\\ &= -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_k\right| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\end{aligned}$$

Decision boundary

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



likelihoods

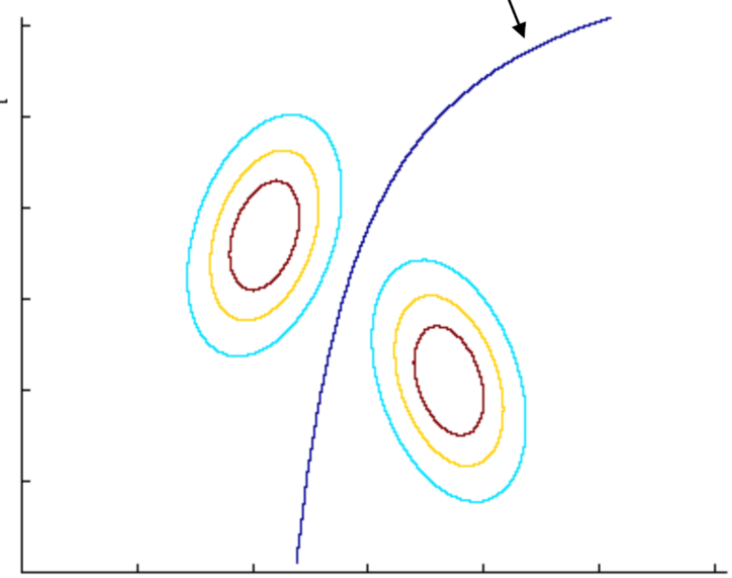


posterior for t_1

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

discriminant:

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



Continued in the next session…