Linear Discriminant Analysis, Naïve Bayes Classifiers

COMP3314 — Lecture 4

Lingpeng Kong
Department of Computer Science, The University of Hong Kong

Based on: Probabilistic Machine Learning by Kevin Murphy
Slides from: Saw Shier Nee with special thanks!

Recap: Supervised Learning in a Nutshell

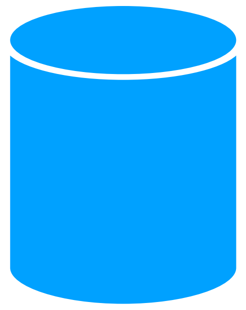
$\boldsymbol{x} \in \mathcal{X}$

inputs/features/covariates/predictors

model

f_θ

training objective/loss



$\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$

training dataset

$\boldsymbol{y} \in \mathcal{Y}$

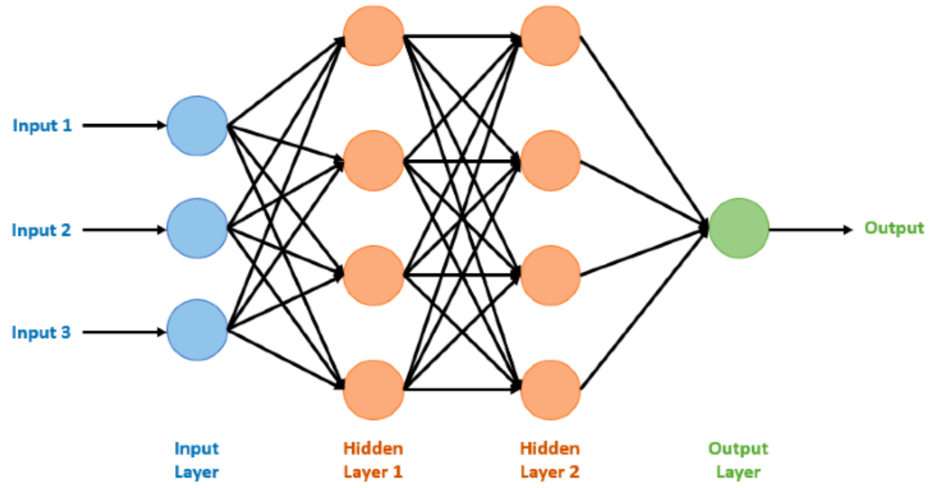
outputs/label/target/response

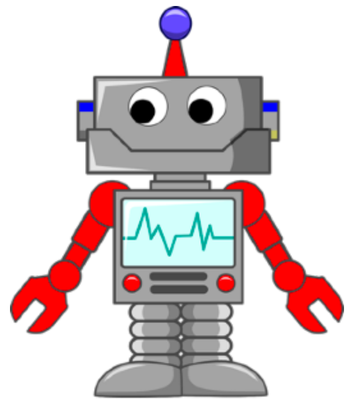
Building a Classifier

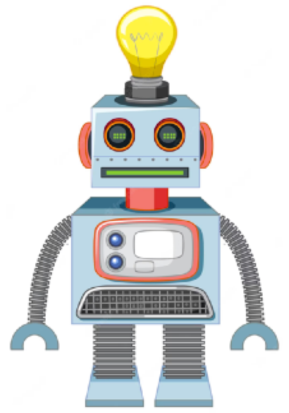
$$p(y = c|\boldsymbol{x};\boldsymbol{\theta}) =$$

for example

$$\sigma(b + w_0 x_0 + w_1 x_1 + \ldots + w_{n-1} x_{n-1} + w_n x_n)$$







$$p(y = c|\boldsymbol{x};\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y = c;\boldsymbol{\theta})\, p(y = c;\boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y = c';\boldsymbol{\theta})\, p(y = c';\boldsymbol{\theta})}$$

On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes

Andrew Y. Ng
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720

Michael I. Jordan
C.S. Div. & Dept. of Stat.
University of California, Berkeley
Berkeley, CA 94720

(Ng and Jordan, 2001)

Bayes' rule

$$p(H = h|Y = y) = \frac{p(H = h)p(Y = y|H = h)}{p(Y = y)}$$

prior distribution — what we know about possible values of H before we see any data

$$p(H)$$

observation distribution — possible outcomes Y we expect to see if H =h

$$p(Y|H = h)$$

likelihood — evaluate the observation distribution at a point corresponding to the actual observations, y

$$p(Y = y|H = h)$$

marginal likelihood

$$p(Y = y) = \sum_{h' \in \mathcal{H}} p(H = h')p(Y = y|H = h') = \sum_{h' \in \mathcal{H}} p(H = h', Y = y)$$

posterior — our new belief state about the possible value of H

$$p(H = h|Y = y)$$

Generative Classifier

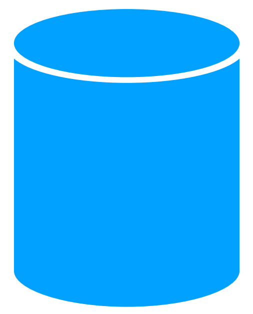
$$p(y = c|\boldsymbol{x}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y = c; \boldsymbol{\theta})p(y = c; \boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y = c'; \boldsymbol{\theta})p(y = c'; \boldsymbol{\theta})}$$

Why the word "generative"?

It specifies a way to generate the features $\boldsymbol{x}$ for each class c, by sampling from $p(\boldsymbol{x} \mid y = c; \boldsymbol{\theta})$

A **discriminative classifier** directly models the class posterior $p(y \mid \boldsymbol{x}; \boldsymbol{\theta})$

Generative Story

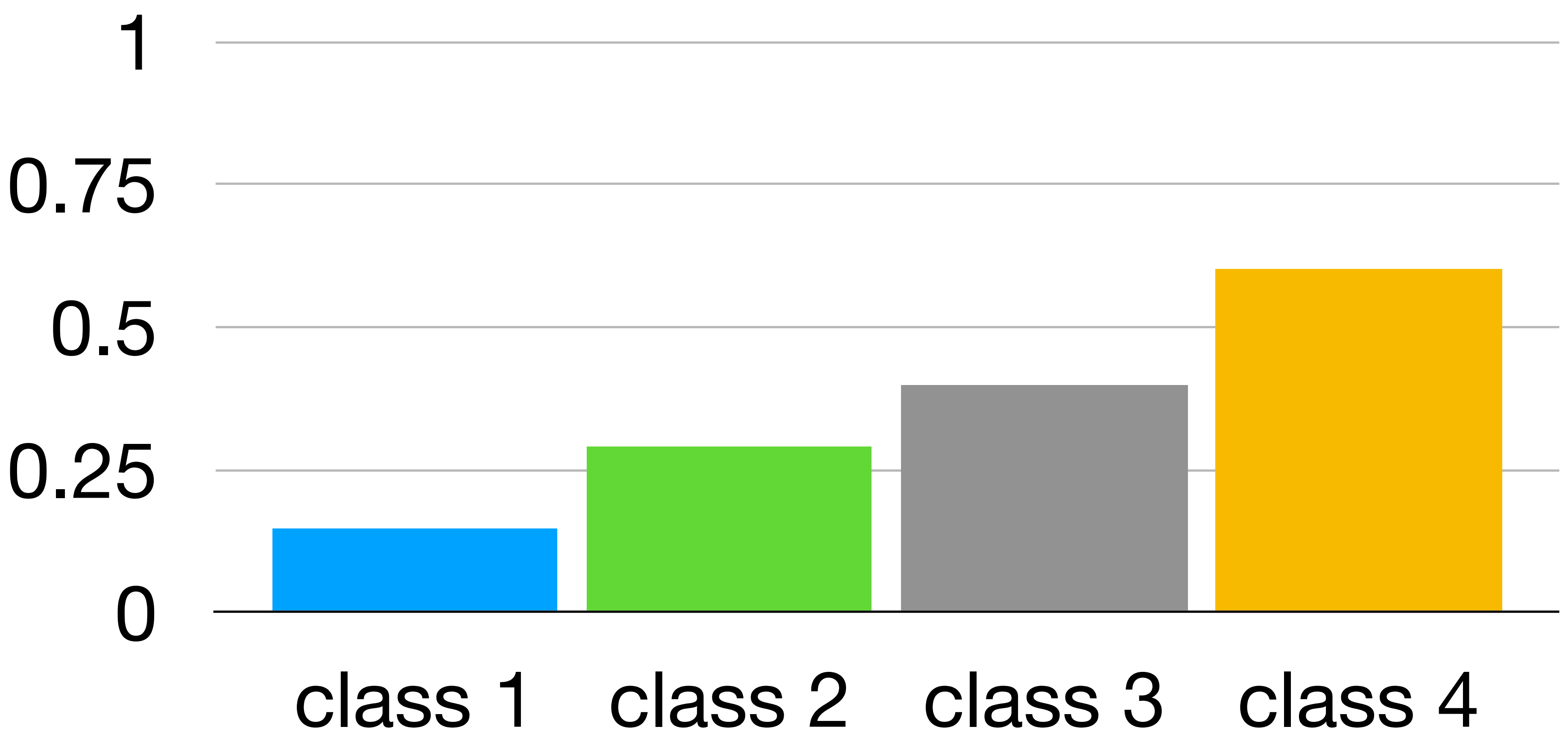


$\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$

training dataset

$$p(y = c|\boldsymbol{x};\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y=c;\boldsymbol{\theta})p(y=c;\boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y=c';\boldsymbol{\theta})p(y=c';\boldsymbol{\theta})}$$

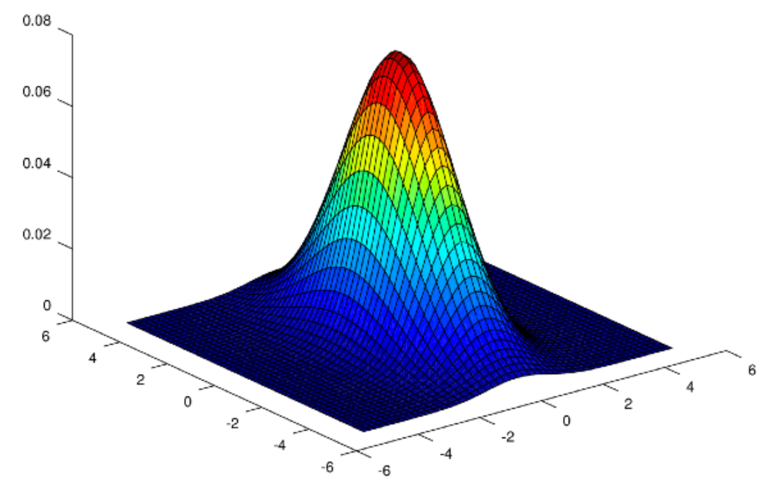
$p(y;\boldsymbol{\theta})$



$\longrightarrow\ y = 3$

$p(\boldsymbol{x} \mid y = 3)$

$\vdots$



$\downarrow$

$x_1 = 1.0 \quad x_2 = 2.3 \quad \cdots$

$(x_1 = 1.0, x_2 = 2.3, \ldots, y = 3)$

Gaussian Discriminant Analysis

$$p(y = c|\boldsymbol{x}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y = c; \boldsymbol{\theta})p(y = c; \boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y = c'; \boldsymbol{\theta})p(y = c'; \boldsymbol{\theta})}$$

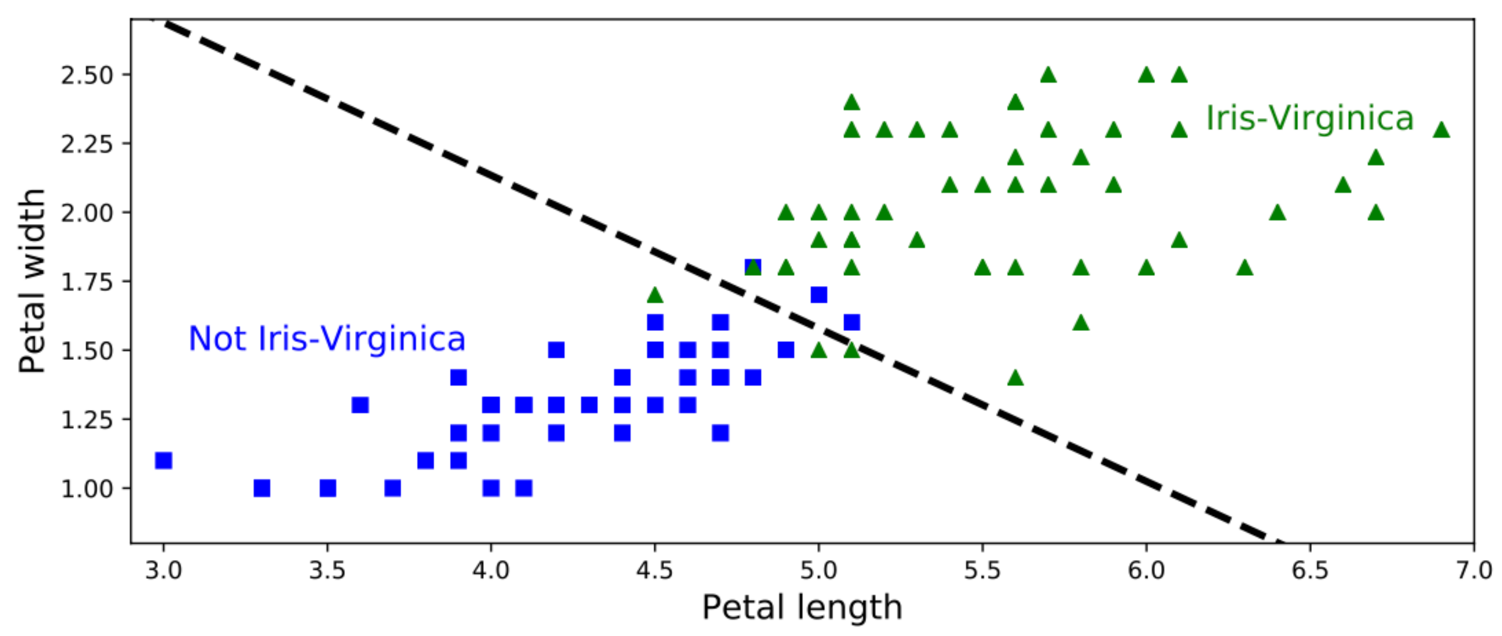
Conditional density: Multivariate Gaussians

$$p(\boldsymbol{x}|y = c, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Gaussian discriminant analysis (GDA):

$$p(y = c|\boldsymbol{x}, \boldsymbol{\theta}) \propto \pi_c \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \qquad \pi_c = p(y = c)$$

Decision Boundaries



Quadratic decision boundaries

$$p(y=c|\boldsymbol{x},\boldsymbol{\theta}) \propto \pi_c \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$$

$$\log p(y=c|\boldsymbol{x},\boldsymbol{\theta}) = \log \pi_c - \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_c| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^\mathsf{T}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) + \text{const}$$

discriminant function

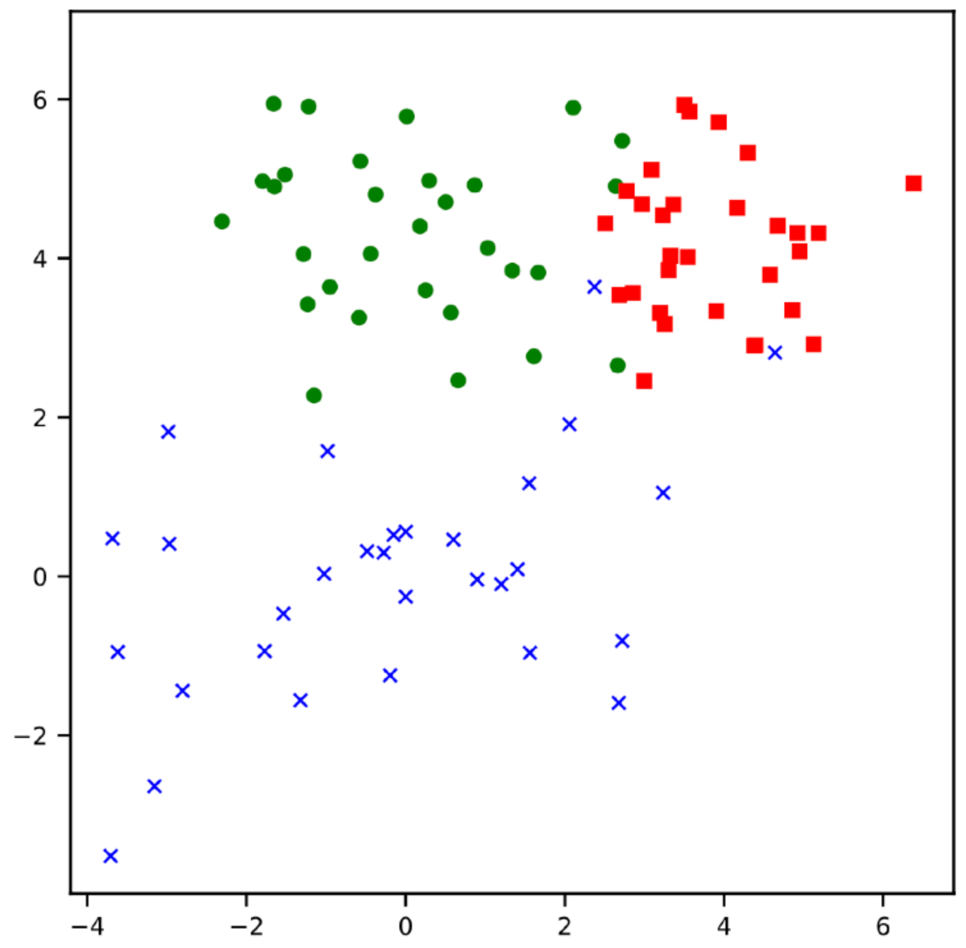
Linear decision boundaries

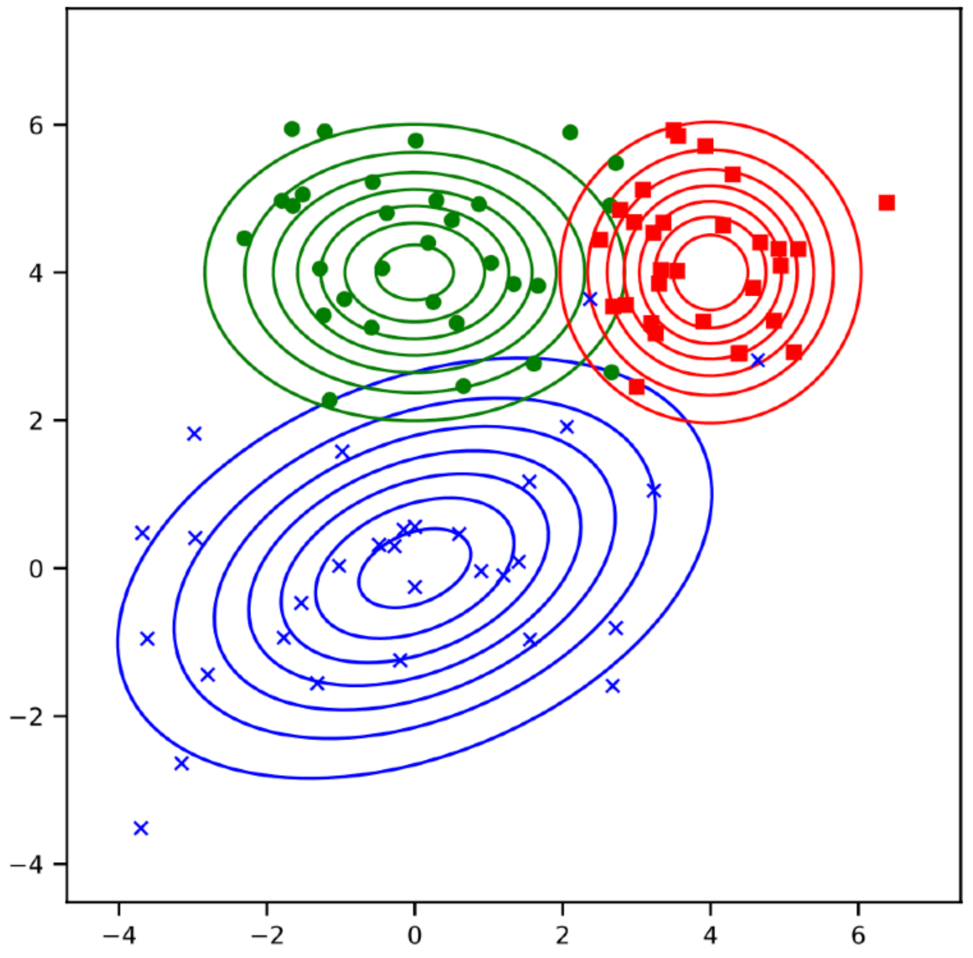
$\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$ 💡 *diagonal LDA — if we further assume a shared diagonal covariance matrix*

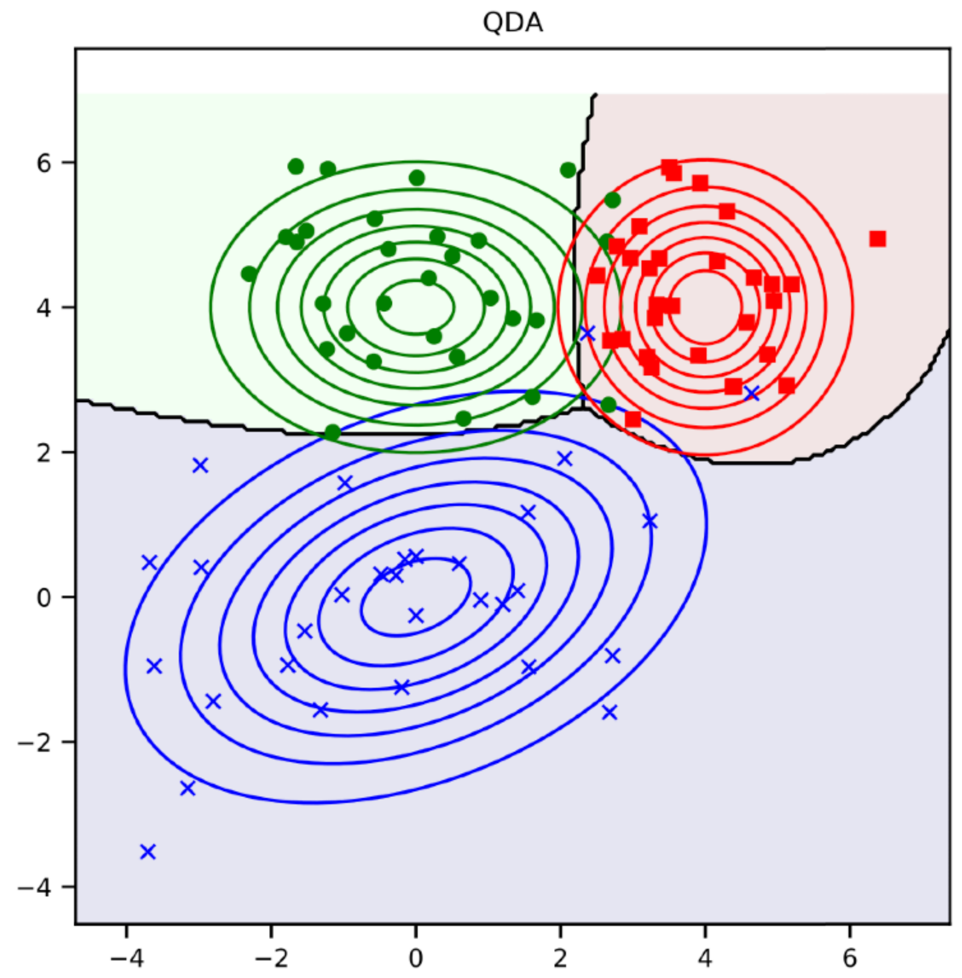
$$\begin{aligned}\log p(y=c|\boldsymbol{x},\boldsymbol{\theta}) &= \log \pi_c - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) + \text{const} \\ &= \underbrace{\log \pi_c - \frac{1}{2}\boldsymbol{\mu}_c^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c}_{\gamma_c} + \boldsymbol{x}^\mathsf{T}\underbrace{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c}_{\boldsymbol{\beta}_c} \underbrace{+\text{const} - \frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}_{\kappa} \\ &= \gamma_c + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_c + \kappa\end{aligned}$$

This is independent of c, hence does not affect our decision!

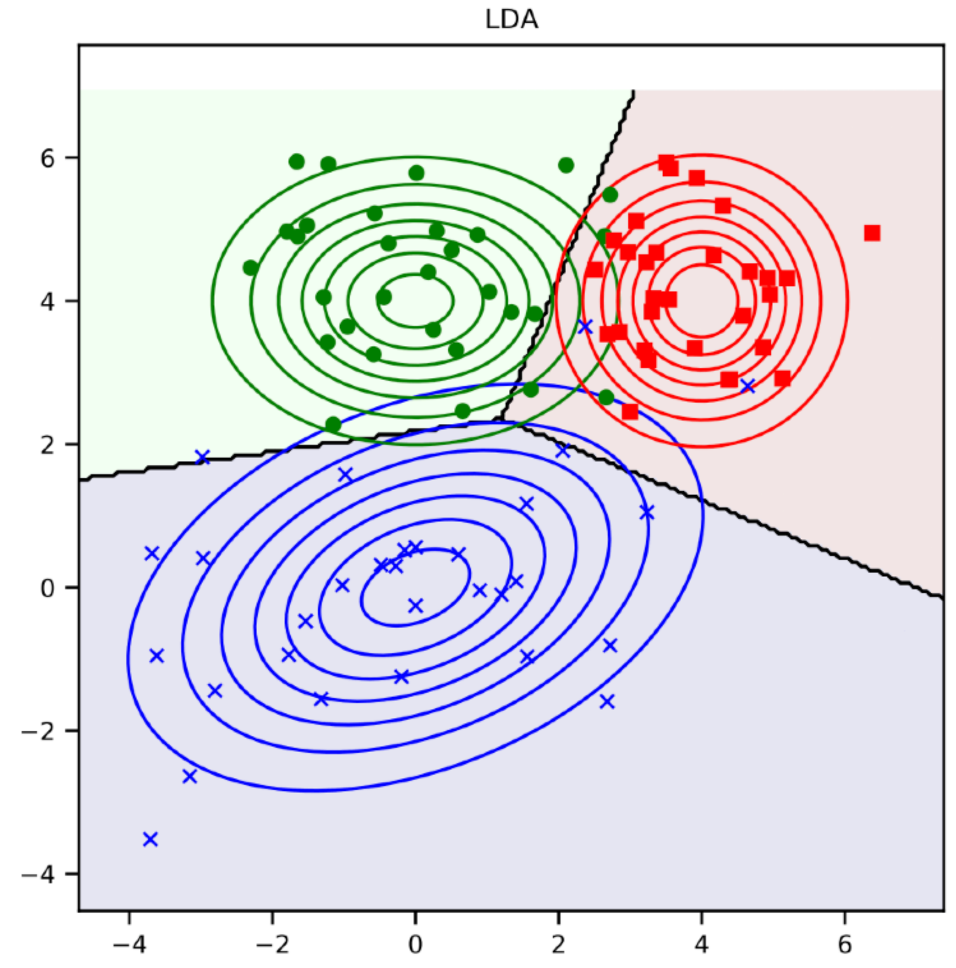
Decision Boundaries







$$\log p(y=c|\boldsymbol{x},\boldsymbol{\theta}) = \log \pi_c - \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_c| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^\mathsf{T}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) + \text{const}$$



$$\log p(y=c|\boldsymbol{x},\boldsymbol{\theta}) \ = \gamma_c + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_c + \kappa$$

Model Fitting (Learning)

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{M}(y_n|\boldsymbol{\pi}) \prod_{c=1}^{C} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)^{\mathbb{I}(y_n=c)}$$

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \left[\sum_{n=1}^{N}\sum_{c=1}^{C} \mathbb{I}\left(y_n = c\right) \log \pi_c\right] + \sum_{c=1}^{C}\left[\sum_{n:y_n=c} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right]$$

MLE: $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{n:y_n=c} \boldsymbol{x}_n$$

$$\hat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c} \sum_{n:y_n=c} (\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_c)(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_c)^{\mathsf{T}}$$

$$\hat{\pi}_c = \frac{N_c}{N}$$

Note: not every parameter has closed form solution

Naïve Bayes Classifier

$$p(y=c|\boldsymbol{x};\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y=c;\boldsymbol{\theta})p(y=c;\boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y=c';\boldsymbol{\theta})p(y=c';\boldsymbol{\theta})}$$

In Gaussian discriminant analysis:

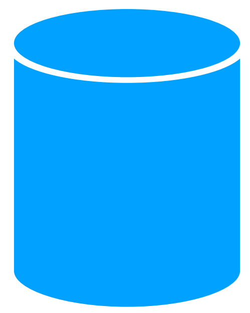
$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

In naïve Bayes classifier:

$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta}) = \prod_{d=1}^{D} p(x_d|y=c,\boldsymbol{\theta}_{dc})$$

This is the naïve part – the naïve Bayes assumption

Generative Story



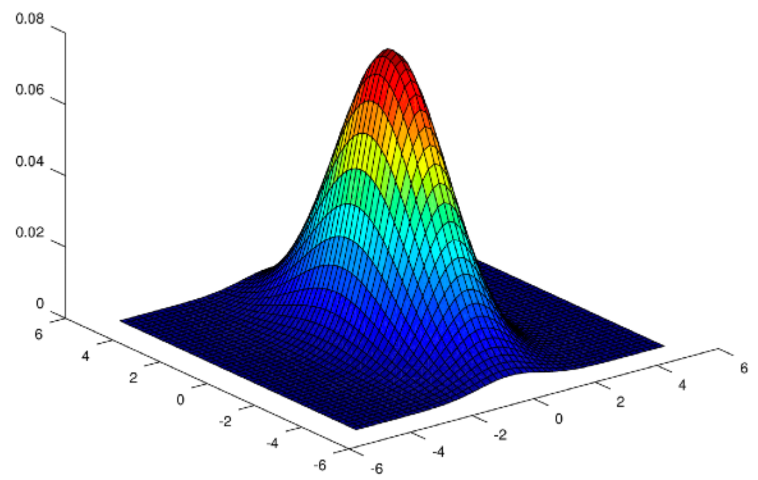
$$\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$$

training dataset

$$p(y = c|\boldsymbol{x}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y = c; \boldsymbol{\theta})p(y = c; \boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y = c'; \boldsymbol{\theta})p(y = c'; \boldsymbol{\theta})}$$

$p(y; \boldsymbol{\theta}) \longrightarrow y = 3$ $\quad p(\boldsymbol{x} \mid y = 3)$

$\vdots$



$\downarrow$

$x_1 = 1.0 \;\; x_2 = 2.3 \;\; \cdots$

$(x_1 = 1.0, x_2 = 2.3, \ldots, y = 3)$

in naïve bayes

$p(x_1 \mid y = 3)$

$p(x_2 \mid y = 3)$

$\vdots$

Naïve Bayes Classifier

	x_1	x_2	x_3	x_4	y
	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	FALSE	No
1	Rainy	Hot	High	TRUE	No
2	Overcast	Hot	High	FALSE	Yes
3	Sunny	Mild	High	FALSE	Yes
4	Sunny	Cool	Normal	FALSE	Yes
5	Sunny	Cool	Normal	TRUE	No
6	Overcast	Cool	Normal	TRUE	Yes
7	Rainy	Mild	High	FALSE	No
8	Rainy	Cool	Normal	FALSE	Yes
9	Sunny	Mild	Normal	FALSE	Yes
10	Rainy	Mild	Normal	TRUE	Yes
11	Overcast	Mild	High	TRUE	Yes
12	Overcast	Hot	Normal	FALSE	Yes
13	Sunny	Mild	High	TRUE	No

Naïve Bayes Classifier

$$p(y=c|\boldsymbol{x},\boldsymbol{\theta}) = \frac{p(y=c|\boldsymbol{\pi})\prod_{d=1}^{D} p(x_d|y=c,\boldsymbol{\theta}_{dc})}{\sum_{c'} p(y=c'|\boldsymbol{\pi})\prod_{d=1}^{D} p(x_d|y=c',\boldsymbol{\theta}_{dc'})}$$

	Outlook (x_1)	Temperature (x_2)	Humidity (x_3)	Windy (x_4)	Play Golf (y)
0	Rainy	Hot	High	FALSE	No
1	Rainy	Hot	High	TRUE	No
2	Overcast	Hot	High	FALSE	Yes
3	Sunny	Mild	High	FALSE	Yes
4	Sunny	Cool	Normal	FALSE	Yes
5	Sunny	Cool	Normal	TRUE	No
6	Overcast	Cool	Normal	TRUE	Yes
7	Rainy	Mild	High	FALSE	No
8	Rainy	Cool	Normal	FALSE	Yes
9	Sunny	Mild	Normal	FALSE	Yes
10	Rainy	Mild	Normal	TRUE	Yes
11	Overcast	Mild	High	TRUE	Yes
12	Overcast	Hot	Normal	FALSE	Yes
13	Sunny	Mild	High	TRUE	No

$p(y = c \mid \boldsymbol{\pi})$

How many parameters?

$$p(y = \text{yes}) = \frac{9}{14}$$

$$p(y = \text{no}) = \frac{5}{14}$$

Naïve Bayes Classifier

$$p(y = c|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{p(y = c|\boldsymbol{\pi}) \boxed{\prod_{d=1}^{D} p(x_d|y = c, \boldsymbol{\theta}_{dc})}}{\sum_{c'} p(y = c'|\boldsymbol{\pi}) \prod_{d=1}^{D} p(x_d|y = c', \boldsymbol{\theta}_{dc'})}$$

	x_1	x_2	x_3	x_4	y
	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	FALSE	No
1	Rainy	Hot	High	TRUE	No
2	Overcast	Hot	High	FALSE	Yes
3	Sunny	Mild	High	FALSE	Yes
4	Sunny	Cool	Normal	FALSE	Yes
5	Sunny	Cool	Normal	TRUE	No
6	Overcast	Cool	Normal	TRUE	Yes
7	Rainy	Mild	High	FALSE	No
8	Rainy	Cool	Normal	FALSE	Yes
9	Sunny	Mild	Normal	FALSE	Yes
10	Rainy	Mild	Normal	TRUE	Yes
11	Overcast	Mild	High	TRUE	Yes
12	Overcast	Hot	Normal	FALSE	Yes
13	Sunny	Mild	High	TRUE	No

$$p(x_d \mid y = c, \boldsymbol{\theta}_{dc})$$

How many parameters?

$$p(x_1 \mid y)$$

	y=yes	y=no	p(x1|y=yes)	p(x2|y=no)
Sunny	3	2	3/9	2/5
Overcast	4	0	4/9	0/5
Rainy	2	3	2/9	3/5

Naïve Bayes Classifier

$$p(y=c|\boldsymbol{x},\boldsymbol{\theta}) = \frac{p(y=c|\boldsymbol{\pi})\prod_{d=1}^{D} p(x_d|y=c,\boldsymbol{\theta}_{dc})}{\sum_{c'} p(y=c'|\boldsymbol{\pi})\prod_{d=1}^{D} p(x_d|y=c',\boldsymbol{\theta}_{dc'})}$$

Outlook

	y=yes	y=no	p(x1|y=yes)	p(x2|y=no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5

Temperature

	y=yes	y=no	p(x1|y=yes)	p(x2|y=no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5

Humidity

	y=yes	y=no	p(x1|y=yes)	p(x2|y=no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5

Wind

	y=yes	y=no	p(x1|y=yes)	p(x2|y=no)
FALSE	6	2	6/9	2/5
TRUE	3	3	3/9	3/5

	y	p(y)
Yes	9	9/14
No	5	5/14

Naïve Bayes Classifier

$$p(y = c|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{p(y = c|\boldsymbol{\pi}) \prod_{d=1}^{D} p(x_d|y = c, \boldsymbol{\theta}_{dc})}{\sum_{c'} p(y = c'|\boldsymbol{\pi}) \prod_{d=1}^{D} p(x_d|y = c', \boldsymbol{\theta}_{dc'})}$$

	y=yes	y=no	p(x1|y=yes)	p(x2|y=no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5

	y=yes	y=no	p(x1|y=yes)	p(x2|y=no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5

	y=yes	y=no	p(x1|y=yes)	p(x2|y=no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5

	y=yes	y=no	p(x1|y=yes)	p(x2|y=no)
FALSE	6	2	6/9	2/5
TRUE	3	3	3/9	3/5

	y	p(y)
Yes	9	9/14
No	5	5/14

$$\text{today} = \{\text{sunny}, \text{hot}, \text{normal}, \text{false}\}$$

$$p(y = \text{yes} \mid \text{today}) = \frac{\frac{9}{14} \cdot \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9}}{p(\text{today})} \quad = 0.67$$

$$p(y = \text{no} \mid \text{today}) = \frac{\frac{5}{14} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5}}{p(\text{today})} \quad = 0.33$$

Naïve Bayes Classifier (Learning)

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{M}(y_n|\boldsymbol{\pi}) \prod_{d=1}^{D} p(x_{nd}|y_n, \boldsymbol{\theta}_d)$$

$$= \prod_{n=1}^{N} \mathcal{M}(y_n|\boldsymbol{\pi}) \prod_{d=1}^{D} \prod_{c=1}^{C} p(x_{nd}|\boldsymbol{\theta}_{d,c})^{\mathbb{I}(y_n=c)}$$

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \left[\sum_{n=1}^{N} \sum_{c=1}^{C} \mathbb{I}\left(y_n = c\right) \log \pi_c\right] + \sum_{c=1}^{C} \sum_{d=1}^{D} \left[\sum_{n:y_n=c} \log p(x_{nd}|\boldsymbol{\theta}_{dc})\right]$$

MLE: $\hat{\pi}_c = \frac{N_c}{N}$

discrete features

$$\hat{\theta}_{dck} = \frac{N_{dck}}{\sum_{k'=1}^{K} N_{dck'}} = \frac{N_{dck}}{N_c}$$

real-value features

$$\hat{\mu}_{dc} = \frac{1}{N_c} \sum_{n:y_n=c} x_{nd}$$

$$\hat{\sigma}^2_{dc} = \frac{1}{N_c} \sum_{n:y_n=c} (x_{nd} - \hat{\mu}_{dc})^2$$