Learning classifiers using generative models

Review: conditional probability

Let $(\Omega, \mathbb{P})$ be a probability space.
(Ω is the sample space; $\mathbb{P}$ is the probability distribution.)

For any events $A, B \subseteq \Omega$,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Bayes' rule:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B \mid A)}{\mathbb{P}(B)}.$$

Let $E, H_0, H_1 \subseteq \Omega$. Conditioned on E, which of H_0 and H_1 is more probable?

Compare $\mathbb{P}(H_0) \cdot \mathbb{P}(E \mid H_0)$ to $\mathbb{P}(H_1) \cdot \mathbb{P}(E \mid H_1)$.

Conditional probability example

Suppose result of test for genetic disease is correct with probability 95%, and suppose the disease is rare: any given person has disease with probability 1%.

Question: If test comes back positive for disease, is it more likely that you have disease or do not?

E = test comes back positive for disease
H_0 = do not have disease
H_1 = have disease
$\mathbb{P}(E \mid H_0) = 0.05$
$\mathbb{P}(E \mid H_1) = 0.95$
$\mathbb{P}(H_0) = 0.99$
$\mathbb{P}(H_1) = 0.01$

Want to compare $\mathbb{P}(H_0 \mid E)$ to $\mathbb{P}(H_1 \mid E)$, so compare

$$\mathbb{P}(H_0) \cdot \mathbb{P}(E \mid H_0) = 0.99 \cdot \underline{0.05} \quad \text{and} \quad \mathbb{P}(H_1) \cdot \mathbb{P}(E \mid H_1) = \underline{0.01} \cdot 0.95.$$

Review: functions of random variables

Let $X \colon \Omega \to \mathcal{X}$ be a random variable.

"Function of a random variable":
For any function $g \colon \mathcal{X} \to \mathbb{R}$,

$$g(X) := g \circ X$$

is also a random variable:

$$g(X)(\omega) = g(X(\omega)).$$

Expected value:

$$\begin{aligned} \mathbb{E}\big(g(X)\big) &= \sum_{\omega \in \Omega} g(X(\omega)) \cdot \mathbb{P}(\omega) \\ &= \sum_{\gamma} \gamma \cdot \mathbb{P}(g(X) = \gamma) \\ &= \sum_{x} g(x) \cdot \mathbb{P}(X = x). \end{aligned}$$

Review: conditional expectation

Let $X \colon \Omega \to \mathcal{X}$ and $Y \colon \Omega \to \mathbb{R}$ be random variables.

Conditional expectation:
For any $x \in \mathcal{X}$ such that $\mathbb{P}(X = x) > 0$:

$$\mathbb{E}[Y \mid X = x] \;:=\; \sum_{y} y \cdot \mathbb{P}(Y = y \mid X = x)\,.$$

What is $\mathbb{E}[Y \mid X]$? A random variable!

$$\mathbb{E}[Y \mid X](\omega) \;=\; \mathbb{E}[Y \mid X = X(\omega)]\,.$$

Law of total expectation:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y \mid X]] \;&=\; \sum_{\omega \in \Omega} \mathbb{E}[Y \mid X = X(\omega)] \cdot \mathbb{P}(\omega) \\ &=\; \sum_{x} \mathbb{E}[Y \mid X = x] \cdot \mathbb{P}(X = x) \;=\; \mathbb{E}(Y)\,.\end{aligned}$$

Bayes classifier (for binary classification)

- Probability distribution P over $\mathcal{X} \times \{0,1\}$; let $(X, Y) \sim P$.
- Think of P as being comprised of two parts.
  1. Marginal distribution of X (a distribution over $\mathcal{X}$).
  2. Conditional distribution of Y given $X = x$, for each $x \in \mathcal{X}$:

$$\eta(x) \;:=\; P(Y = 1 \mid X = x)\,.$$

- The optimal classifier with smallest error rate (i.e., *Bayes classifier*) is

$$f^\star(x) \;=\; \begin{cases} 0 & \text{if } \eta(x) \leq 1/2 \\ 1 & \text{if } \eta(x) > 1/2\,. \end{cases}$$

(Formal derivation on next slide.)

Formal derivation

Error rate of $f \colon \mathcal{X} \to \{0,1\}$ can be written as

$$\mathrm{err}_P(f) \;=\; P(f(X) \neq Y) \;=\; \mathbb{E}\big[\mathbb{1}\{f(X) \neq Y\}\big]\,.$$

Define $g \colon \mathcal{X} \to \mathbb{R}$ by

$$g(x) \;:=\; \mathbb{E}\big[\mathbb{1}\{f(X) \neq Y\} \mid X = x\big]\,.$$

Then, by law of total probability,

$$\begin{aligned} g(x) \;&=\; P(Y = 0 \mid X = x) \cdot \mathbb{1}\{f(x) \neq 0\} + P(Y = 1 \mid X = x) \cdot \mathbb{1}\{f(x) \neq 1\} \\ &=\; (1 - \eta(x)) \cdot \mathbb{1}\{f(x) = 1\} + \eta(x) \cdot \mathbb{1}\{f(x) = 0\}\,. \end{aligned}$$

What should $f(x)$ be so that $g(x)$ is as small as possible?

Since $f(x) \in \{0,1\}$, best to have

$$f(x) \;:=\; \begin{cases} 0 & \text{if } \eta(x) \leq 1 - \eta(x) \\ 1 & \text{if } \eta(x) > 1 - \eta(x) \end{cases}$$

. . . which is the same as $f^\star(x)$. $\square$

Bayes classifier (for K-class classification)

- Probability distribution P over $\mathcal{X} \times \{1, 2, \dots, K\}$; let $(X, Y) \sim P$.
- Think of P as being comprised of two parts.
  1. Marginal distribution of X (a distribution over $\mathcal{X}$).
  2. Conditional distribution of Y given $X = x$, for each $x \in \mathcal{X}$.
- Bayes classifier:

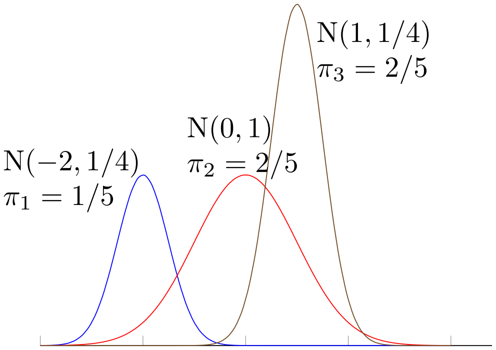
$$f^\star(x) \;=\; \operatorname*{arg\,max}_{y \in \{1,2,\dots,K\}} P(Y = y \mid X = x)\,.$$

Structure of Bayes classifier

- By Bayes' rule, Bayes classifier can be written as

$$f^*(x) \;=\; \operatorname*{arg\,max}_{y\in\{1,2,\dots,K\}} \; P(Y=y)\cdot P(X=x \mid Y=y)\,.$$

- Motivates thinking of P as being comprised of:
  1. *Class priors* $\pi_1, \pi_2, \dots, \pi_K \in [0,1]$, where $\pi_y = P(Y=y)$.
  2. *Class conditional distributions* $P_1, P_2, \dots, P_K$, where P_y is conditional distribution of X given $Y=y$.



Generative models

- In context of classification problems, a *generative model* is a statistical model $\mathcal{P}$ on $\mathcal{X} \times \{1, 2, \dots, K\}$, where each $P_{\boldsymbol{\theta}} \in \mathcal{P}$ is

$$P_{\boldsymbol{\theta}}(x, y) \;=\; \pi_y \cdot P_{t_y}(x)\,,$$

where the parameters are

$$\boldsymbol{\theta} \;=\; (\pi_1, \pi_2, \dots, \pi_K, t_1, t_2, \dots, t_K)\,.$$

- Typically, all class conditional distributions $P_{t_1}, P_{t_2}, \dots, P_{t_K}$ come from same statistical model (e.g., Gaussian distribution family).
- Form of Bayes classifier corresponding to $P_{\boldsymbol{\theta}}$:

$$x \;\mapsto\; \operatorname*{arg\,max}_{y\in\{1,2,\dots,K\}} \; \pi_y \cdot P_{t_y}(x)\,.$$

Learning a classifier

Basic approach to learning a classifier using a generative model

Suppose we observe data $D = \{(x_i, y_i)\}_{i=1}^n$, regarded as an i.i.d. sample.

0. Partition $\{x_i\}_{i=1}^n$ into $D_1, D_2, \dots, D_K$, where

$$D_y \;=\; \{x_i : y_i = y\}\,.$$

1. Estimate $\pi_1, \pi_2, \dots, \pi_K$ using $\{y_i\}_{i=1}^n$ (e.g., using MLE: $\hat{\pi}_y := |D_y|/n$).
   For each $y \in \{1, 2, \dots, K\}$: estimate t_y using D_y (e.g., using MLE).
2. Let classifier $\hat{f}$ be the Bayes classifier for distribution $P_{\widehat{\boldsymbol{\theta}}}$ corresponding to parameter estimates

$$\widehat{\boldsymbol{\theta}} \;=\; (\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_K, \hat{t}_1, \hat{t}_2, \dots, \hat{t}_K)\,,$$

i.e.,

$$\hat{f}(x) \;=\; \operatorname*{arg\,max}_{y\in\{1,2,\dots,K\}} \; \hat{\pi}_y \cdot P_{\hat{t}_y}(x)\,.$$

Example: Gaussian class conditional densities

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, and using Gaussian class conditional densities.

Data $D = \{(x_i, y_i)\}_{i=1}^n$, regarded as an i.i.d. sample.

0. Split D into D_0, D_1, where $D_0 := \{x_i : y_i = 0\}$ and $D_1 := D \setminus D_0$.
1. Estimate parameters:

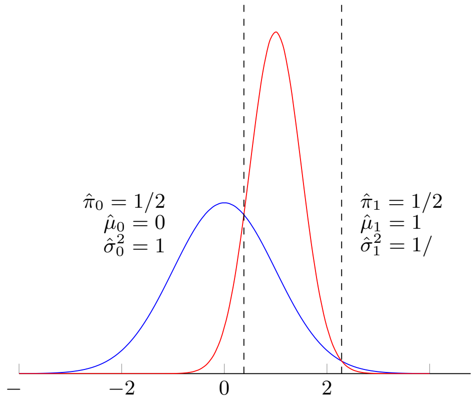
$$\begin{aligned} \hat{\pi}_0 &:= |D_0|/n\,, & \hat{\pi}_1 &:= |D_1|/n\,, \\ \hat{\mu}_0 &:= \text{sample mean}(D_0)\,, & \hat{\mu}_1 &:= \text{sample mean}(D_1)\,, \\ \hat{\sigma}_0^2 &:= \text{sample variance}(D_0)\,, & \hat{\sigma}_1^2 &:= \text{sample variance}(D_1)\,. \end{aligned}$$

2. Form classifier $\hat{f}$ using

$$\hat{f}(x) \;=\; \operatorname*{arg\,max}_{y\in\{0,1\}} \hat{\pi}_y \cdot \varphi_{\hat{\mu}_y, \hat{\sigma}_y^2}(x)$$

where φ_{μ,σ^2} is the $\mathrm{N}(\mu, \sigma^2)$ density.

Example: Gaussian class conditional densities



Classifier:

$$\hat{f}(x) = \begin{cases} 1 & \text{if } x \in [0.38, 2.29]; \\ 0 & \text{otherwise.} \end{cases}$$

Dotted lines = *decision boundary*.

Bayes classifiers

Suppose $\{(x_i, y_i)\}_{i=1}^n$ is an iid sample from P, and let $\mathcal{P}$ be a generative model with parameter space Θ.

- Let $\hat{\boldsymbol{\theta}} \in \Theta$ be parameter estimate obtained using data $\{(x_i, y_i)\}_{i=1}^n$.
- Let $\hat{f}$ be the Bayes classifier for distribution $P_{\hat{\boldsymbol{\theta}}} \in \mathcal{P}$.

Let $f^\star$ be the Bayes classifier for P.

- Is $\hat{f}$ the same as $f^\star$?

High-dimensional feature spaces

Suppose $\mathcal{X} = \{0,1\}^d$ or $\mathcal{X} = \mathbb{Z}_+^d$ or $\mathcal{X} = \mathbb{R}^d$ where $d > 1$.

What statistical models can we use for the class conditional distributions?

- *non-parametric models*: very general, but quality may be poor for large d.
- Often have prior knowledge about *statistical dependencies* between variables. Leverage this knowledge to form a *graphical model*.
- Some simple models: *multivariate Gaussians*, *product distributions*.

Product distributions on $\{0,1\}^d$

Suppose $\mathcal{X} = \{0,1\}^d$, and let $\mathcal{P}$ be all *product distributions on* $\{0,1\}^d$.

Parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_d) \in [0,1]^d$:

$$P_{\boldsymbol{\mu}}(\boldsymbol{x}) = \prod_{j=1}^{d} \mu_j^{x_j} (1 - \mu_j)^{1 - x_j} \quad \text{for all } \boldsymbol{x} = (x_1, x_2, \dots, x_d) \in \{0,1\}^d .$$

If **random vector** $\boldsymbol{X} = (X_1, X_2, \dots, X_d) \sim P_{\boldsymbol{\mu}}$, then $X_1, X_2, \dots, X_d$ are independent random variables, and

$$P_{\boldsymbol{\mu}}(X_j = 1) = \mu_j .$$

Naïve Bayes classifiers

Generative models that use product distributions as class conditionals
⟶ *Naïve Bayes classifiers.*

(Using product distributions on $\{0,1\}^d$ as in previous slide.)
What is the form of Bayes classifier corresponding to distribution with parameters $\boldsymbol{\theta} = (\pi_1, \pi_2, \ldots, \pi_K, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K)$?

$$\begin{aligned}
\boldsymbol{x} \quad \mapsto \quad & \underset{y\in\{1,2,\ldots,K\}}{\arg\max} \ \pi_y \cdot P_{\boldsymbol{\mu}_y}(\boldsymbol{x}) \\
= \quad & \underset{y\in\{1,2,\ldots,K\}}{\arg\max} \ \log\Bigl(\pi_y \cdot P_{\boldsymbol{\mu}_y}(\boldsymbol{x})\Bigr) \\
= \quad & \underset{y\in\{1,2,\ldots,K\}}{\arg\max} \ \log\left(\pi_y \cdot \prod_{j=1}^{d} {\mu_{y,j}}^{x_j} (1-\mu_{y,j})^{1-x_j}\right) \\
= \quad & \underset{y\in\{1,2,\ldots,K\}}{\arg\max} \ \log \pi_y + \sum_{j=1}^{d} x_j \log \mu_{y,j} + (1-x_j)\log(1-\mu_{y,j}) \\
= \quad & \underset{y\in\{1,2,\ldots,K\}}{\arg\max} \ b_y + \langle \boldsymbol{w}_y, \boldsymbol{x} \rangle
\end{aligned}$$

for some appropriate definition of $b_y \in \mathbb{R}$ and $\boldsymbol{w}_y \in \mathbb{R}^d$ in terms of π_y and $\boldsymbol{\mu}_y$.

Standard Gaussian distributions on $\mathbb{R}^d$

Standard normal (Gaussian) distribution on $\mathbb{R}^1$

$X \sim \mathrm{N}(0,1)$, density

$$\varphi_{0,1}(x) \ = \ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad \text{for all } x \in \mathbb{R}\,.$$

Standard normal (Gaussian) distribution on $\mathbb{R}^d$

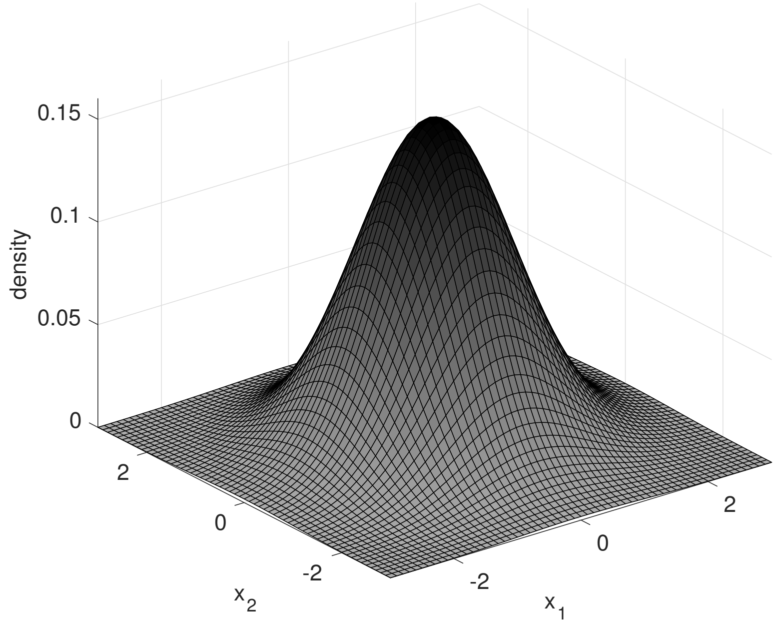
$\boldsymbol{X} = (X_1, X_2, \ldots, X_d) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$, density

$$\varphi_{\boldsymbol{0},\boldsymbol{I}}(\boldsymbol{x}) \ = \ \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) \quad \text{for all } \boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d\,.$$

Usually written as

$$\varphi_{\boldsymbol{0},\boldsymbol{I}}(\boldsymbol{x}) \ = \ \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\boldsymbol{x}\|_2^2}{2}\right).$$

Standard normal (Gaussian) distribution on $\mathbb{R}^d$



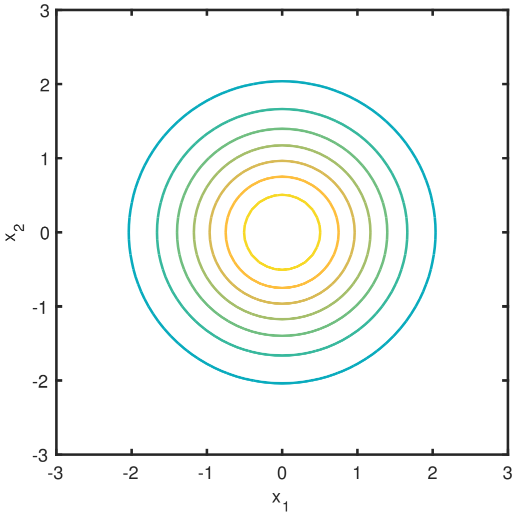
Standard normal (Gaussian) distribution on $\mathbb{R}^d$

- $\mathbb{E}(X_i) = 0$
- $\mathrm{var}(X_i) = \mathrm{cov}(X_i, X_i) = 1$
- $\mathrm{cov}(X_i, X_j) = 0$ for $i \neq j$.

Arrange means into a vector and covariances in a $d{\times}d$ matrix:

$$\mathbb{E}(\boldsymbol{X}) \ = \ \boldsymbol{0}\,, \quad \mathrm{cov}(\boldsymbol{X}) \ = \ \boldsymbol{I}$$

(zero vector and identity matrix).



Contours of equal standard normal density in $\mathbb{R}^2$

General Gaussian distributions on $\mathbb{R}^d$

(General) Gaussian distributions on $\mathbb{R}^d$ come from applying two operations to another (e.g., the standard) Gaussian distribution:

$$\overbrace{\boldsymbol{x} \;\mapsto\; \mathrlap{\underbrace{\phantom{\boldsymbol{A}\boldsymbol{x} \;\mapsto\; \boldsymbol{A}\boldsymbol{x}+\boldsymbol{\mu}}}_{\text{translation}}}\boldsymbol{A}\boldsymbol{x}}^{\text{linear map}} \;\mapsto\; \boldsymbol{A}\boldsymbol{x}+\boldsymbol{\mu}$$

for some vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and *invertible* linear map $\boldsymbol{A} \in \mathbb{R}^{d\times d}$.

Fact: Let $\boldsymbol{\mu} \in \mathbb{R}^d$ be any vector, and $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be any invertible matrix. For any random vector $\boldsymbol{X}$ in $\mathbb{R}^d$, the random vector $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}$ satisfies

$$\mathbb{E}(\boldsymbol{Y}) \;=\; \boldsymbol{\mu}\,, \quad \operatorname{cov}(\boldsymbol{Y}) \;=\; \boldsymbol{A}\boldsymbol{A}^\top\,.$$

Furthermore, if $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{A}^\top)$.

Examples of linear maps

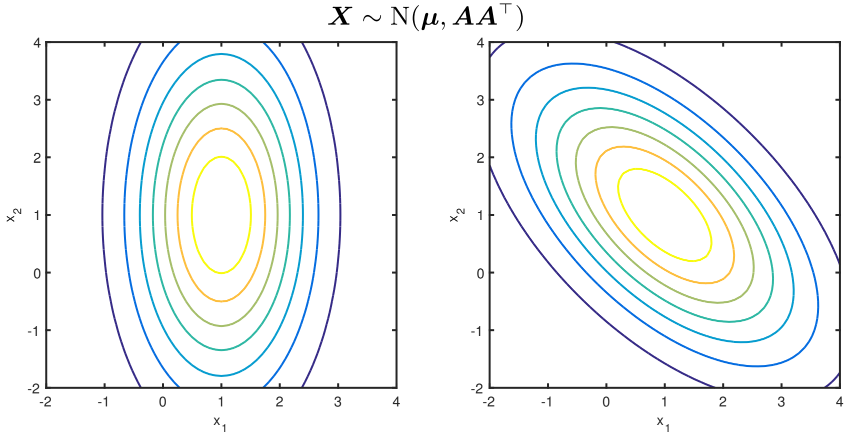
Write $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

1. If $\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, then $\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} x_1 \\ 2x_2 \end{bmatrix}$.

   (Scale coordinates x_1 and x_2 by, respectively, 1 and 2.)
2. If $\boldsymbol{A} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, then $\boldsymbol{A}\boldsymbol{x} = x_1 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} + 2x_2 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$.

   (Coordinate scaling as above, followed by rotation.)

General Gaussian distributions on $\mathbb{R}^d$



$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\boldsymbol{A} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

Gaussian parametric model

Gaussian parametric model:
$\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^d \times \mathbb{S}_{++}$, with

$$\mathbb{S}_{++} \;=\; \text{all symmetric positive definite (p.d.) } d{\times}d \text{ real matrices}\,.$$

Recall: a matrix $\boldsymbol{M}$ is p.d. if and only if

$$\boldsymbol{v}^\top \boldsymbol{M}\boldsymbol{v} > 0 \quad \text{for all } \boldsymbol{v} \neq \boldsymbol{0}\,.$$

Example: Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be invertible. For any $\boldsymbol{v} \neq \boldsymbol{0}$,

$$\begin{aligned} \boldsymbol{v}^\top(\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{v} \;&=\; (\boldsymbol{A}^\top\boldsymbol{v})^\top(\boldsymbol{A}^\top\boldsymbol{v}) \\ &=\; \boldsymbol{w}^\top\boldsymbol{w} \qquad (\text{letting } \boldsymbol{w} := (\boldsymbol{A}^\top)^{-1}\boldsymbol{v}) \\ &=\; \|\boldsymbol{w}\|_2^2 \;\geq\; 0\,. \end{aligned}$$

The only vector with zero length is $\boldsymbol{0}$, so $\|\boldsymbol{w}\|_2^2 = 0$ if and only if $\boldsymbol{w} = \boldsymbol{0}$. Furthemore, $\boldsymbol{w} = \boldsymbol{0}$ if and only if $\boldsymbol{v} = \boldsymbol{A}^\top\boldsymbol{w} = \boldsymbol{0}$. But $\boldsymbol{v} \neq \boldsymbol{0}$, so $\boldsymbol{w} \neq \boldsymbol{0}$ and $\|\boldsymbol{w}\|_2^2 > 0$. Hence, $\boldsymbol{A}\boldsymbol{A}^\top$ is p.d.

MLE for Gaussian parameters

- Suppose we observe $\{\boldsymbol{x}_i\}_{i=1}^n$ (regarded as an i.i.d. sample). What is the MLE for the Gaussian parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$?
- Log-likelihood of $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given $\{\boldsymbol{x}_i\}_{i=1}^n$:

$$\log \mathcal{L}(\boldsymbol{\theta}; \{\boldsymbol{x}_i\}_{i=1}^n) = \sum_{i=1}^n \left[\log \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right) \right].$$

- Gradient of above with respect to $\boldsymbol{\mu}$:

$$-\sum_{i=1}^n \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}).$$

- Lo and behold, MLE for $\boldsymbol{\mu}$ is the *sample mean*:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i .$$

- Using *matrix derivatives* gives the MLE for $\boldsymbol{\Sigma}$, the *sample covariance*:

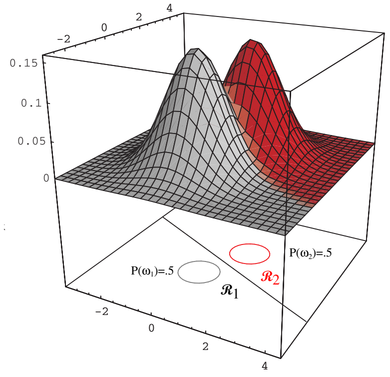
$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^\top$$

(where $\hat{\boldsymbol{\mu}}$ is the sample mean). **What could go wrong?**

Multivariate Gaussian class conditionals

Example: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$, and using multivariate Gaussian class conditional densities.

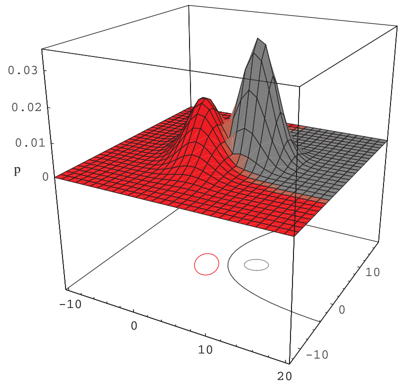
Bayes classifier corresponding to distribution with parameters $\boldsymbol{\theta} = (\pi_0, \pi_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$:



$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$

Bayes classifier:

linear decision boundary



$\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$

Bayes classifier:

quadratic decision boundary

Generative models with parameter tying

Sometimes, we must estimate (some of) the parameters of the class conditional distributions *jointly*, rather than separately.

Example: multivariate Gaussian class conditionals *with shared covariance*:

$$t_1 = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad t_2 = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), \quad \ldots, \quad t_K = (\boldsymbol{\mu}_K, \boldsymbol{\Sigma}).$$

This is called *parameter tying*.

MLEs for $\pi_1, \pi_2, \ldots, \pi_K, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$ given $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$:

$$\hat{\pi}_y := |D_y|/n, \qquad \hat{\boldsymbol{\mu}}_y := \text{sample mean}(D_y).$$

But what's the MLE for the shared class conditional covariance $\boldsymbol{\Sigma}$?

$$\widehat{\boldsymbol{\Sigma}} := \frac{1}{n} \sum_{y=1}^K \sum_{\boldsymbol{x}_i \in D_y} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_y)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_y)^\top .$$

Final remarks

Some redeeming qualities of classifiers based on generative models:

- Simple recipe, many variations.
- Can leverage domain knowledge about class conditional distributions.
- Can be very efficient when K is large.

Critical drawbacks:

- Classifier relies on formula (via Bayes' rule) that assumes the estimated class priors and conditional distributions are perfect, which is not true.
- Modeling P away from decision boundary between classes is wasted effort: not necessary for good classification.

Key takeaways

1. Generative structure of Bayes classifier.
2. Basic properties of multivariate Gaussians.
3. Basic recipe for learning a classifier based on a generative model, and concept of parameter tying.
4. Specific generative models with product distributions and multivariate Gaussians as class conditionals.
5. High-level advantages and disadvantages of classifiers based on generative models.