

TD 1 — Deux gaussiennes : erreur de Bayes, seuils et métriques

Séance 1

Christophe Ambroise

2026-09-29

Objectifs

Calculer une erreur de Bayes à la main sur un modèle où tout est connu, puis retrouver les mêmes quantités sur un échantillon simulé. Comparer les deux.

Tout le TD porte sur le même modèle, celui du cours :

$$X | Y = 0 \sim \mathcal{N}(\mu_0, \sigma^2), \quad X | Y = 1 \sim \mathcal{N}(\mu_1, \sigma^2), \quad P(Y = 1) = \pi_1,$$

avec $\mu_0 = -1$, $\mu_1 = 1,5$, $\sigma = 1$ et $\pi_1 = 0,35$. La classe 1 est celle des **positifs** — les malades d'un dépistage. Une seule variable, des variances égales : tout se calcule en forme close.

Partie théorique

Exercice 1 — Probabilité a posteriori et seuil de décision

1. Écrire $P(Y = 1 | X = x)$ à l'aide de π_0 , π_1 , f_0 et f_1 .
2. Montrer que cette probabilité s'écrit $\sigma(ax + b)$, où $\sigma(u) = 1/(1 + e^{-u})$, et donner a et b .
3. En déduire que décider « positif » lorsque $P(Y = 1 | X = x) > s$ revient à décider « positif » lorsque $x > t$, pour un seuil t que l'on exprimera en fonction de s . Que deviendrait cette équivalence si $\mu_1 < \mu_0$?

Exercice 2 — Seuil optimal, risque de Bayes, coûts asymétriques

1. Exprimer le risque sous coût 0-1 d'une règle « décider 1 si $x > t$ » :

$$R(t) = \pi_0 P(X > t | Y = 0) + \pi_1 P(X \leq t | Y = 1).$$

L'écrire avec la fonction de répartition Φ de la loi normale centrée réduite.

2. Minimiser $R(t)$ et retrouver le seuil de Bayes t^* . Le calculer numériquement.
3. Un faux négatif coûte c fois un faux positif. Reprendre le calcul et donner t_c . Dans quel sens la frontière se déplace-t-elle quand c croît ?
4. Calculer $R(t)$ pour trois seuils : t^* , le seuil $t = 0,25$ obtenu en ignorant les proportions a priori, et t_3 (faux négatif trois fois plus coûteux).

Partie empirique

Exercice 3 — Simuler l'échantillon et voir les densités

Le code de simulation et d'affichage est fourni. L'exécuter, puis répondre.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

MU0, MU1, SIGMA, PI1 = -1.0, 1.5, 1.0, 0.35
PI0 = 1 - PI1

rng = np.random.default_rng(0)
n = 2000
y = (rng.random(n) < PI1).astype(int)      # étiquettes
x = rng.normal(np.where(y == 1, MU1, MU0), SIGMA)  # observations

grille = np.linspace(-5, 6, 400)
plt.hist(x[y == 0], bins=40, density=True, alpha=0.25,
         color="#41597a", label="échantillon, classe 0")
plt.hist(x[y == 1], bins=40, density=True, alpha=0.25,
         color="#d96125", label="échantillon, classe 1")
plt.plot(grille, PI0 * norm.pdf(grille, MU0, SIGMA), color="#41597a", lw=2,
         label=r"$\pi_0 f_0$")
plt.plot(grille, PI1 * norm.pdf(grille, MU1, SIGMA), color="#d96125", lw=2,
         label=r"$\pi_1 f_1$")
plt.axvline(0.498, color="black", ls="--", label="seuil $t^*$")
plt.legend(); plt.xlabel("x"); plt.ylabel("densité")
plt.show()
```

```
mu0 <- -1.0; mu1 <- 1.5; sigma <- 1.0; pi1 <- 0.35; pi0 <- 1 - pi1

set.seed(0)
n <- 2000
y <- rbinom(n, 1, pi1)      # étiquettes
x <- rnorm(n, ifelse(y == 1, mu1, mu0), sigma)  # observations

grille <- seq(-5, 6, length.out = 400)
hist(x[y == 0], breaks = 40, freq = FALSE, col = "#41597a40", border = NA,
     xlim = c(-5, 6), xlab = "x", ylab = "densité", main = "")
hist(x[y == 1], breaks = 40, freq = FALSE, col = "#d9612540", border = NA,
     add = TRUE)
lines(grille, pi0 * dnorm(grille, mu0, sigma), col = "#41597a", lwd = 2)
lines(grille, pi1 * dnorm(grille, mu1, sigma), col = "#d96125", lwd = 2)
abline(v = 0.498, lty = 2)
legend("topright", bty = "n", lwd = 2, col = c("#41597a", "#d96125"),
     legend = c(expression(pi[0] * f[0]), expression(pi[1] * f[1])))
```

1. Les histogrammes sont normalisés séparément par classe ; les courbes tracées sont $\pi_k f_k$. Pourquoi les deux ne se superposent-ils pas ? Comment corriger l'affichage pour qu'ils coïncident ?
2. Vérifier sur l'échantillon la proportion de positifs. À quelle distance de $\pi_1 = 0,35$ est-elle ? Cet écart est-il compatible avec $n = 2000$?

3. Repérer sur la figure la zone qui correspond au risque de Bayes.

Exercice 4 — Matrices de confusion aux trois seuils

Pour chacun des trois seuils de l'exercice 2, construire la matrice de confusion sur l'échantillon simulé, puis calculer l'erreur, la précision et le rappel.

1. Écrire le code (une décision est simplement $\mathbf{x} > \mathbf{t}$).
2. Comparer l'erreur empirique au $R(t)$ théorique, seuil par seuil.
3. Que devient la matrice de confusion du classifieur constant « toujours négatif » ? Quelle est son exactitude ? Son rappel ?

Exercice 5 — Courbe ROC et AUC

1. Donner les expressions théoriques de $\text{TPR}(t)$ et $\text{FPR}(t)$, et tracer la courbe ROC théorique en faisant varier t .
2. Construire la courbe ROC empirique à partir de l'échantillon, et la superposer à la théorique.
3. Montrer que l'AUC théorique vaut $\Phi\left(\frac{\mu_1 - \mu_0}{\sigma\sqrt{2}}\right)$, et la comparer à l'AUC empirique.
4. Placer les trois seuils de l'exercice 2 sur la courbe. Où se trouve le seuil optimal ? Est-ce le point le plus proche du coin supérieur gauche ?

Exercice 6 — Score F_1 et déséquilibre des classes

1. Écrire la précision au niveau de la population : $\text{précision}(t) = \frac{\pi_1 \text{TPR}(t)}{\pi_1 \text{TPR}(t) + \pi_0 \text{FPR}(t)}$, puis $F_1(t)$.
2. Tracer $F_1(t)$ et l'erreur $R(t)$ sur le même axe des seuils. Les deux sont-elles optimales au même endroit ?
3. Refaire le calcul avec $\pi_1 = 0,02$, les moyennes inchangées. Que deviennent l'AUC, la précision au seuil t^* , et le F_1 ?
4. Conclure : quand publier une ROC, quand publier une précision-rappel ?

Remarque finale

! À retenir

Sur ce modèle, seuil optimal, risque de Bayes, ROC et AUC se calculent. L'échantillon les retrouve à $1/\sqrt{n}$ près. Sur données réelles, seule la colonne « empirique » existe.

Les deux scripts exécutables correspondants sont `python/td01_deux_gaussiennes.py` et `r/td01_deux_gaussiennes.R`.