Exercício 1 — Naive Bayes na mão

Objetivos de aprendizagem

- Calcular predições de um Naive Bayes manualmente a partir de uma tabela categórica.
- Entender como incorporar features contínuas ao Naive Bayes.

Considere o seguinte dataset de frutas, com o target `Banana` indicando se cada registro se refere a uma banana:

ID	Cor	Forma	Origem	Banana
1	amarela	alongada	importada	sim
2	amarela	redonda	local	não
3	amarela	alongada	importada	não
4	marrom	alongada	importada	sim
5	marrom	redonda	local	não
6	verde	redonda	importada	sim
7	verde	alongada	local	não
8	vermelha	redonda	importada	não

Queremos usar um classificador Naive Bayes para prever se uma nova fruta é banana.

(a) Estime a posterior $\hat{\pi}(\mathbf{x}_*)$ para a nova observação $\mathbf{x}_* = (\text{amarela}, \text{redonda}, \text{importada})$ e diga como você classificaria esse objeto. Mostre as contas passo a passo.

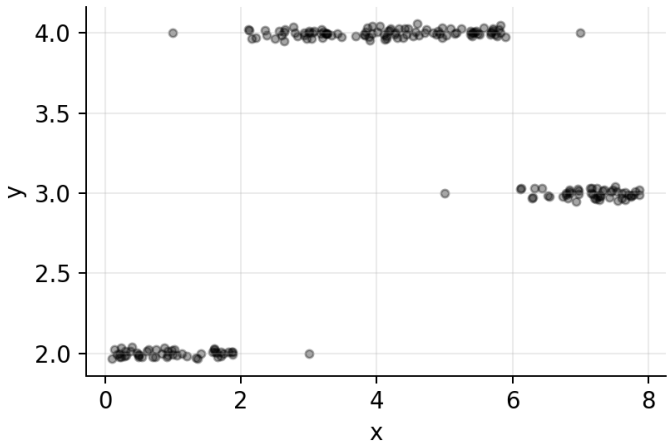
(b) Suponha que, além dos três atributos categóricos, você passa a dispor de uma nova feature `Comprimento`, medida em cm (valor numérico contínuo). Descreva, em uma ou duas frases, como incorporar essa feature ao Naive Bayes.

Exercício 2 — Análise discriminante (LDA e QDA)

Objetivos de aprendizagem

- Configurar uma análise discriminante manualmente a partir de dados visuais.
- Fazer predições com análise discriminante.
- Discutir as diferenças entre LDA e QDA.

A figura abaixo mostra um dataset $\mathcal{D}$ com $n = 200$ observações, onde x é uma feature contínua unidimensional e y um target contínuo. Queremos prever y a partir de x, mas primeiro vamos transformá-lo em um problema de classificação.



Defina a nova variável discreta z a partir de y segundo

$$z^{(i)} = \begin{cases} 1, & y^{(i)} \in (-\infty,\, 2{,}5] \\ 2, & y^{(i)} \in (2{,}5,\, 3{,}5] \\ 3, & y^{(i)} \in (3{,}5,\, \infty). \end{cases}$$

Vamos agora aplicar *quadratic discriminant analysis* (QDA) nesses dados.

(a) Estime visualmente, a partir da figura, as médias condicionais $\mu_k = \mathbb{E}(x \mid z = k)$ para cada classe $k \in \{1, 2, 3\}$. Uma estimativa grosseira basta.

(b) Esboce (à mão ou em um gráfico) as densidades estimadas por classe, $\hat{p}(x \mid z = k)$, que o QDA produziria para esses dados. Justifique em uma ou duas frases o formato das curvas escolhido.

(c) Como o esboço do item (b) mudaria se usássemos LDA (*linear discriminant analysis*) em vez de QDA? Explique a diferença.

(d) Dadas duas novas observações $x_{*1} = -10$ e $x_{*2} = 7$, e supondo classes aproximadamente do mesmo tamanho, qual seria a predição do QDA para cada uma? Como você chegou a ela?

(e) Por que o QDA é preferível ao LDA nesses dados?

Exercício 3 — Fronteiras de decisão no dataset `cassini`

Objetivos de aprendizagem

- Implementar em Python LDA, QDA e Naive Bayes com `scikit-learn`.
- Comparar visualmente as fronteiras de decisão produzidas por cada classificador.

Neste exercício você vai treinar três classificadores clássicos — LDA, QDA e Naive Bayes — no dataset sintético `cassini` (três classes, duas features) e comparar suas fronteiras de decisão.

Baixe o CSV em https://raw.githubusercontent.com/slds-lmu/lecture_i2ml/master/exercises/data/cassini_data.csv. Ele tem três colunas: `x.1`, `x.2` e `classes` (rótulo). Use o snippet abaixo para carregá-lo e perturbar a dimensão `x.2` com ruído gaussiano de média zero e desvio padrão 0,5:

```
import numpy as np
import pandas as pd

URL = ("https://raw.githubusercontent.com/slds-lmu/lecture_i2ml/"
       "master/exercises/data/cassini_data.csv")
df = pd.read_csv(URL)

rng = np.random.default_rng(123)
df["x.2"] = df["x.2"] + rng.normal(0, 0.5, size=len(df))

X = df[["x.1", "x.2"]].to_numpy()
y = df["classes"].to_numpy()
```

(a) Faça um scatter plot dos dados (após a perturbação), colorindo os pontos pela classe. Descreva em uma frase o formato das três classes.

(b) Treine os três classificadores usando `sklearn`: `LinearDiscriminantAnalysis`, `QuadraticDiscriminantAnal` e `GaussianNB`.

(c) Para cada classificador, desenhe a fronteira de decisão sobre uma grade 2D cobrindo o intervalo dos dados (uma forma simples é avaliar `model.predict` em todos os pontos da grade e usar `contourf`). Apresente os três gráficos lado a lado, com os pontos de treino por cima.

(d) Compare as fronteiras. Em até dois parágrafos, discuta: qual classificador acompanha melhor o formato das classes? Onde o LDA falha? Há diferença visível entre QDA e Naive Bayes?

Dica

Aqui você está avaliando os modelos no próprio conjunto de treino — esses gráficos servem para comparar o formato das fronteiras de decisão, não para estimar erro de generalização. A parte de avaliação honesta (train/test, cross-validation) é assunto da próxima lista.