

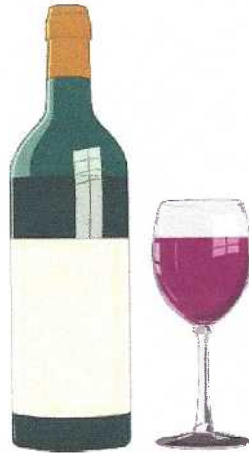
Generative modeling in one dimension

Topics we'll cover

- ① Generative modeling at work
- ② The Gaussian in one dimension

A classification problem

You have a bottle of wine whose label is missing.



Which winery is it from, 1, 2, or 3?

Solve this problem using visual and chemical features of the wine.

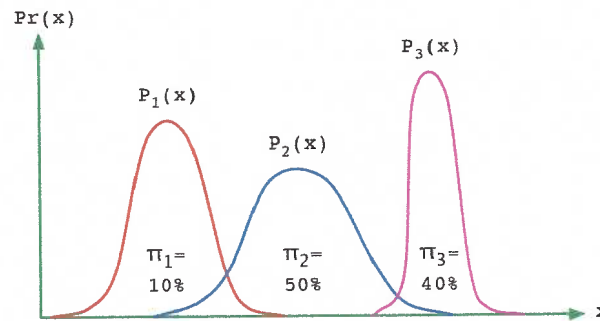
The data set

Training set obtained from 130 bottles

- Winery 1: 43 bottles
- Winery 2: 51 bottles
- Winery 3: 36 bottles
- For each bottle, 13 features:
 - 'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash', 'Magnesium',
 - 'Total phenols', 'Flavanoids', 'Nonflavanoid phenols', 'Proanthocyanins',
 - 'Color intensity', 'Hue', 'OD280/OD315 of diluted wines', 'Proline'

Also, a separate test set of 48 labeled points.

Recall: the generative approach



$$\begin{aligned} \pi_1 P_1(x) \\ \pi_2 P_2(x) \\ \pi_3 P_3(x) \end{aligned}$$

For any data point $x \in \mathcal{X}$ and any candidate label j ,

$$\Pr(y = j|x) = \frac{\Pr(y = j)\Pr(x|y = j)}{\Pr(x)} = \frac{\pi_j P_j(x)}{\Pr(x)}$$

Optimal prediction: the class j with largest $\pi_j P_j(x)$.

↓ Probability that the label is j or $y=j$ given features x

Fitting a generative model

Training set of 130 bottles:

- Winery 1: 43 bottles, winery 2: 51 bottles, winery 3: 36 bottles
- For each bottle, 13 features: 'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash', 'Magnesium', 'Total phenols', 'Flavanoids', 'Nonflavanoid phenols', 'Proanthocyanins', 'Color intensity', 'Hue', 'OD280/OD315 of diluted wines', 'Proline'

Class weights:

$$\pi_1 = 43/130 = 0.33, \quad \pi_2 = 51/130 = 0.39, \quad \pi_3 = 36/130 = 0.28$$

Need distributions P_1, P_2, P_3 , one per class.

Base these on a single feature: 'Alcohol'.

To simplify, instead of 13 features, let's just use 1 → alcohol

$$\begin{array}{r} 43 \\ 51 \\ + 36 \\ \hline 130 \text{ bottles} \end{array}$$

$$\begin{array}{r} 1 \text{ } 2 \\ 0.33 \\ 0.39 \\ 0.28 \\ \hline 1.00 \end{array}$$

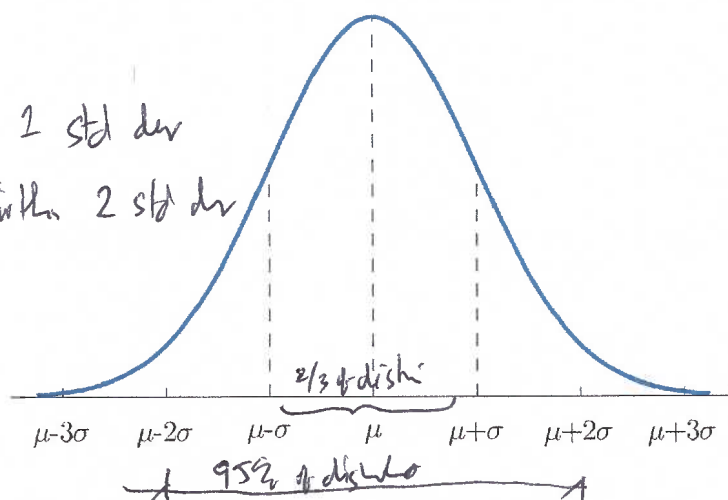
add up to 1

The univariate Gaussian

$\frac{2}{3}$ dishes ^{lies} within 1 std dev
95% dishes lies within 2 std dev

centered at mean μ
Symmetric about the mean

Mean μ
Std dev σ
 $N(\mu, \sigma^2)$

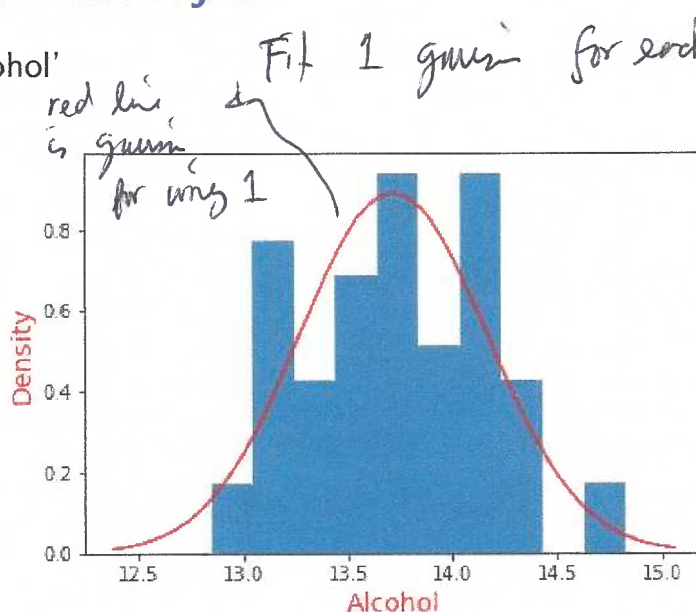


The Gaussian $N(\mu, \sigma^2)$ has mean μ , variance σ^2 , and density function

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The distribution for winery 1

Single feature: 'Alcohol'



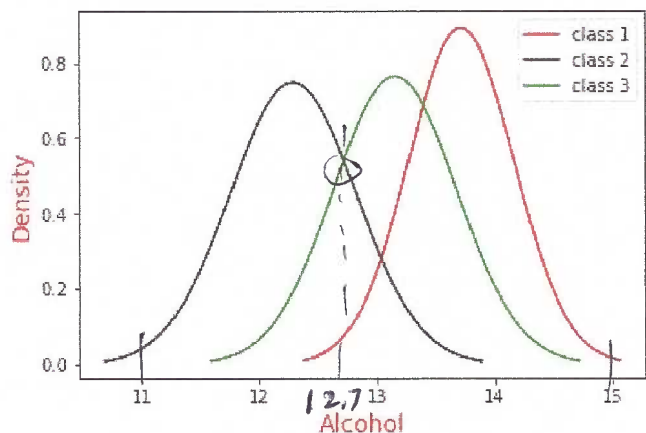
TO FIT A GAUSSIAN,

we just need to figure out the mean & variance
for winery 1, we take the 43 number the 43 alcohol levels & compute mean & variance

Mean $\mu = 13.72$, Standard deviation $\sigma = 0.44$ (variance 0.20)

All three wineries

For one feature alone,
distributions are overlapping so classifying might not be accurate or not useful
practically error



- $\pi_1 = 0.33, P_1 = N(13.7, 0.20)$ ←

- $\pi_2 = 0.39, P_2 = N(12.3, 0.28)$

- $\pi_3 = 0.28, P_3 = N(13.2, 0.27)$

for 12.7, we would pick
either black or green

To classify x : Pick the j with highest $\pi_j P_j(x)$

Test error: $14/48 = 29\%$

error rate for one
feature - alcohol