2.2.3.1 Bayes Classification

In an ideal world, we have for each class $c = 1, \dots, C$ the probability for x to belong to the c-th class.
For example if we are classifying fruits, K_1 could be "Banana", K_2 could be "Apple" etc
When we write $\mathbb{P}(K_c|x)$ we are asking: "Given that we observed x, what is the probability that it belongs to class c?"
If we denote the event of belonging to class c by K_c, we know the conditional probabilities $\mathbb{P}(x|K_c)$ (in English: *conditional distributions*), e.g. $\mathbb{P}(1003g, \text{red}|\text{Banana}) = 0\%$. Thus, we can use Bayes' theorem to obtain

$$\mathbb{P}(K_c|x) = \frac{\mathbb{P}(K_c)\mathbb{P}(x|K_c)}{\mathbb{P}(x)} = \frac{\mathbb{P}(K_c)\mathbb{P}(x|K_c)}{\sum_{i=1}^{C} \mathbb{P}(K_i)\mathbb{P}(x|K_i)}$$

Explanation

- $\mathbb{P}(K_c)$ is the **prior probability** of class c, i.e. how common this class is in general. It is denoted as π_c
- $\mathbb{P}(x|K_c)$ is the **likelihood**, meaning the probability of observing x given that it belongs to class c
- $\mathbb{P}(x)$ is the **total probability** of x over all possible classes, which is computed as $\sum_{i=1}^{C} \mathbb{P}(K_i)\mathbb{P}(x|K_i)$

The probability $\mathbb{P}(K_c)$ of the c-th class is usually abbreviated as π_c (in English: *class prior*), leading to

2.2.9

$$\mathbb{P}(K_c|x) = \frac{\pi_c \mathbb{P}(x|K_c)}{\sum_{i=1}^{C} \pi_c \mathbb{P}(x|K_i)}$$

Classification is then done by choosing the most probable class, i.e.

$$\arg\max_{c=1}^{C} \frac{\pi_c \mathbb{P}(x|K_c)}{\sum_{i=1}^{C} \pi_c \mathbb{P}(x|K_i)} = \arg\max_{c=1}^{C} \pi_c \mathbb{P}(x|K_c)$$

(Since the denominator is the same for all classes, we can just compute the numerator!)
This means:

- Compute $\pi_c \mathbb{P}(x|K_c)$ for each class
- Pick the class with the highest value!

Example: Classifying fruits

Imagine we want to classify a fruit x based on **weight** and **color**. Suppose we have

- Bananas (K_1): 90% of bananas are yellow and 10% are green
- Apples (K_2): 50% are red, 40% are green and 10% are yellow
- Cherries (K_3): 95% are red, 5% are yellow

If we observe $x = (1003g, red)$, the probability of observing that for each fruit is

- $\mathbb{P}(x|K_1)$ (prob of a 1003g red banana) is 0% --> bananas are never red
- $\mathbb{P}(x|K_2)$ is 50%
- $\mathbb{P}(x|K_3)$ is 95%
  If apples and cherries are equally common, we classify x as a **cherry** since it has the highest probability

Back to the script

The mapping $x \mapsto \arg\max_{c=1}^{C} \pi_c \mathbb{P}(x|K_c)$ is also called the Bayes classifier.
This means that given some feature x, we assign it to the class K_c that maximizes $\pi_c \mathbb{P}(x|K_c)$
Note that this is independent of the denominator in 2.2.9
Often, x is a continuous random variable, making it meaningless to work with $\mathbb{P}(x|K_c)$: If $x|K_c$ is, for example, normally distributed, then $\mathbb{P}(x|K_c) = 0$ for all $x \in \mathbb{R}^d$
In this case, however we can use Bayes' theorem for densities and obtain

2.2.10

$$\mathbb{P}(K_c|x) = \frac{\pi_c p(x|K_c)}{\sum_{i=1}^{C} \pi_c p(x|K_i)}$$

where $p(x|K_c)$ are the conditional **densities** ($p(x|K_c)$ now is a PDF of x given class K_c). Note that the left side still represents a probability and not a density since the possible classes form a discrete and finite probability space!
In practe, we unfortunately have neither π_i nor $\mathbb{P}(x|K_c)$ or $p(x|K_c)$ available.
Both quantities must be approximated using the data. To this end, one can use the class proportions:

$$\hat{\pi}_c := \frac{\#\{i = 1, \ldots, N \mid y_i = l_c\}}{N}$$

This just means:

- l_c represents the label for the class
- $y_i = l_c$ means "the i-th sample is in class c"

- Count how many times class c appears in the training data
- Divide by the total number of samples N
  ... and is used regardless of discrete or continuous case (see below!)

Additionally, the conditional probabilities $\mathbb{P}(x|K_c)$ or densities $p(x|K_c)$ must be suitably approximated by $\hat{P}(x|K_c)$ or $\hat{p}(x|K_c)$.

- $\hat{P}(x|K_c)$ is an approximation of the probability $\mathbb{P}(x|K_c)$ when x is discrete (e.g. categorical features)
- $\hat{p}(x|K_c)$ is an approximation of the probability density function $p(x|K_c)$ when x is continuous (e.g. real-valued features)

We then obtain the approximate Bayes classifier:

$$x \mapsto \arg\max_{c=1}^{C} \hat{\pi}\hat{P}(x|K_c) \quad \text{or} \quad x \mapsto \arg\max_{c=1}^{C} \hat{\pi}_c\hat{p}(x|K_c)$$

This means:

1. Compute the estimated prior π_c+
2. Compute the estimated probability (depending on the case discrete or continuous)
3. Assign x to the class c that maximizes

Discrete Random Variables

If $x|K_c$ is a discrete random variable (e.g. colors red, green, yellow), one can approximate $\mathbb{P}(x|K_c)$ by proportions in the training data, e.g.

$$\hat{P}(\text{red}|K_c) := \frac{\#\{i|x_i = \text{red}, y_i = l_c\}}{\#\{i|y_i = l_c\}}$$

This means:

- **Numerator**: number of samples where x_i (the feature) is red **and** the label y_i is class c
- **Denominator**: total number of samples in class c

Example

Sample	Color (x_i)	Fruit Class (y_i)
1	Red	Apple
2	Green	Apple
3	Yellow	Banana
4	Red	Apple
5	Yellow	Banana
6	Red	Cherry
7	Green	Apple
8	Yellow	Banana
9	Red	Apple

There are 3 red apples and 4 apples in total so

$$\hat{P}(\text{red}|K_{\text{apple}}) = \frac{3}{4} = 0.75$$

This means that 75% of apples in our training data are red

Gaussian Mixture Model

For continuous variables, a common assumption is that $x|K_c \sim \mathcal{N}(\mu_c, \sigma_c^2)$ is normally distributed with means μ_c and variances σ_c^2 for $c = 1, \dots, C$. For simplicity, we will only consider the one-dimensional case here.
We can approximate the densities $p(x|K_c)$ for $x \in \mathbb{R}$ by

$$\hat{p}(x|K_c) := \frac{1}{\sqrt{2\pi\hat{\sigma_c}^2}} \exp\left(-\frac{(x - \hat{\mu}_c)^2}{2\hat{\sigma_c}^2}\right)$$

for $c = 1, \dots, C$ where

$$\mu_c = \frac{1}{\#\{i|y_i = l_c\}} \sum_{i|y_i = l_c} x_i \qquad \hat{\sigma}_c^2 := \frac{1}{\#\{i|y_i = l_c\}} \sum_{i|y_i = l_c} (x_i - \mu_c)^2$$

are the mean and discrete variance of the points with label l_c

- μ_c is just the average of all feature values x_i that belong to class c
  - The denominator is the number of training samples in class c
  - The numerator sums all x_i values for class c
- $\hat{\sigma_c}^2$ is the variance estimate, which calculates the **spread** of values around the meaen

- Each sample's difference from the mean is squared and summed, then divided by the number of samples of training points in class c

Kernel Density Classification

If the distribution of $x|K_c$ is more complicated, the Gaussian Mixture Model may not be a good choice. In this case, one can approximate it using a so-called kernel density estimator.
We also restrict ourselves here to the case $d = 1$, i.e. the inputs are **one-dimensional**.
We define

2.2.12

$$\hat{p}(x|K_c) := \frac{1}{\#\{i|y_i = l_c\}} \sum_{i|y_i=l_c} K_\lambda(x - x_i)$$

where $\lambda > 0$ is the bandwidth, $K_\lambda(x) := \frac{1}{\lambda}K(\frac{x}{\lambda})$ and $K : \mathbb{R} \to \mathbb{R}$ is a non-negative function with $\int_{\mathbb{R}} K(x)dx = 1$. Possible choices are the
Gaussian kernel

$$K(x) := (2\pi)^{-1/2}\exp(-x^2/2)$$

or $K(x) := \frac{1}{2}1_{|x|\leq 1}$. However note that for the second choice, it may happen that $\hat{p}(x|K_c) = 0$ for all $c = 1, \ldots, C$ which does not allow for class assignment.
This approach depends on the choice of bandwidth λ.
In particular, for $0 < \lambda \ll 1$, $\hat{p}(x|K_c) \approx 0$ if x is not a training data point x_i.
On the other hand, if λ is very large, $\hat{p}(\cdot|K_c)$ is nearly constant.
It can be shown that an optimal choice of λ for large values of $n \in \mathbb{N}$ takes the form $\lambda = Cn^{-\frac{1}{5}}$.
For normally distributed data $x_i \sim \mathcal{N}(\mu, \sigma^2)$, the optimal constant is given by $C = (\frac{4}{3})^{\frac{1}{5}}\sigma$