IAML - Week 1

Antonio León Villares

September 2021

Contents

1 Maths and Probability

- Probability allows us to deal with uncertainty in ML models
- Bayes' Theorem allows us to derive conditional probabilities from other conditional probabilities
- We use the Maximum Likelihood Approach to estimate a distribution, by considering which distribution is most likely to have produced some observed data

1.1 Probability in ML

- **What is probability?**
  - branch of maths concerned with **describing** the likelihood of events
  - allows us to numerically manipulate and understand **uncertainty**
  - uncertainty depends on what we know
- **Why is probability used in machine learning?**
  - training data can have uncertainty (i.e sensors are unreliable)
  - algorithms can be analysed or developed via probability theory

1.2 Random Variables

- **What is the sample space?**
  - *Experiment*: a procedure that can be infinitely repeated, with a well-defined set of outcomes
  - *Sample Space*: the set of all possible outcomes of an *experiment*
  - *Event*: a subset of the *sample space*, to which probabilities are assigned
- **What do random variables represent?**
  - *Random Variable*: a variable whose possible values are numerical outcomes of a random phenomenon
  - in other words, a *random variable* maps the sample space to a set of states, each of which corresponds to one event
  - the set of states is **mutually exclusive** and **collectively exhaustive**
  - ***Example***: if we toss a coin, let X be a *random variable*. Then, we can assign $X = 0$ to the outcome "heads", and $X = 1$ to the outcome "tails"

- **What defines a discrete probability distribution?**
  - a *probability mass function*, which gives the probability of a **discrete random variable** taking any particular value:

$$P(X = x)$$

  - for a DRV X and a PMF $P(X = x)$, we expect:

$$\sum_x P(X = x) = 1$$

  - ***Example***: if we consider the roll of a die, any face is equally likely to be the outcome of a throw. There are 6 sides, so:

$$P(X = x) = \frac{1}{6}$$

    is the PMF defining the outcomes of an experiment in which a die is rolled

- **What defines a continuous probability distribution?**
  - a *probability density function*, $p(x)$, which gives the probability of a **continuous random variable** taking some interval of values (if X is a CRV, then $P(X = x) = 0, \forall x$), such as:

$$P(a \leq X \leq b)$$

  - to calculate probabilities with PDFs, we use integration:

$$P(a \leq X \leq b) = \int_a^b p(x)dx$$

  - we require a PDF which is always positive ($p(x) > 0$), and:

$$\int p(x)dx = 1$$

  - ***Example***: a **uniform distribution**. Consider a CRV X, defined on the interval $[0, N]$. Then, its PDF is:

$$p(x) = \begin{cases} \frac{1}{N}, & x \in [0, N] \\ 0, & otherwise \end{cases}$$

    We can check that:

$$\int_0^N \frac{1}{N}dx = \left[\frac{x}{N}\right]_0^N = 1$$

- **What is the mean and variance of a distribution?**
  - *Expected Value*: the typical/average value that a RV takes
  - *Variance*: a measure of how far away the values of RVs deviate from their mean (expected value)
  - for a RV X, its expected value is $\mu_X = E(X)$ and its variance is $\sigma_X^2 = Var(X)$
  - for a DRV:

$$E(X) = \sum_x x \times P(X = x)$$

$$\begin{aligned} Var(X) &= E[(X - E(X))^2] \\ &= E(X^2) - [E(X)]^2 \\ &= \sum_x (x - E(X))^2 P(X = x) \\ &= \left(\sum_x x^2 P(X = x) \right) - [E(X)]^2 \end{aligned}$$

1.3 Joint Distributions

- **What does a joint distribution represent?**
  - considers the outcome of 2 events occurring at the same time
  - if X and Y are random variables, then their joint probability distribution is given by:

$$P(X = x, Y = y)$$

    and can be expressed by using a table:

	X = *sunny*	X = *rainy*
Y = *ice cream*	0.7	0.05
Y = *hot cocoa*	0.01	0.24

Table 1: The sum of all probabilities must be exactly 1

- **How can you calculate $P(X = x)$ using joint distributions?**
  - this is called a *marginal probability*:

$$P(X = x) = \sum_y P(X = x, Y = y)$$

    Thus:

$$P(Y = \text{ice cream}) = P(X = \text{sunny}, Y = \text{ice cream}) + P(X = \text{rainy}, Y = \text{ice cream}) = 0.75$$

1.4 Conditional Probability and Bayes' Rule

- **How is conditional dependence between RVs defined?**
  - we can consider the effect of an event happening on the probability of another event taking place
  - this is encompassed by conditional probability:

$$P(X = x|Y = y)$$

  - to calculate conditional probability:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

  - in general:

$$P(X_1, X_2, \ldots, X_n|Y) = \frac{P(X_1, X_2, \ldots X_n, Y)}{P(Y)}$$

- **What are the product and chain rules in probability?**
  - the product rule allows us to compute full probabilities, based on conditional probabilities:

$$P(X, Y) = P(Y) \times P(X|Y) = P(X) \times P(Y|X)$$

  - more generally, we get the chain rule:

$$\begin{aligned} P(X_1, X_2, \ldots X_n, Y) &= P(X_1|X_2, \ldots X_n, Y) \times P(X_2, \ldots X_n, Y) \\ &= \prod_{i=1}^{n} P(X_i|X_{i+1}, \ldots, X_n, Y) \end{aligned}$$

- **What does Bayes' Theorem state?**
  - Bayes' Theorem allows us to express conditional probabilities in terms of each other:

$$P(Y|X) = \frac{P(Y) \times P(X|Y)}{P(X)}$$

  - if $P(X)$ is not known directly, then we can use marginal probabilities:

$$P(X) = \sum_y P(X, Y) = \sum_y P(X|Y) \times P(Y)$$

- **What are the prior and posterior probabilities in the Bayes Theorem formulation?**

- if we consider Y to be our current beliefs, and X some newly observed data, then $P(Y|X)$ can be thought as our uncertainty about our current beliefs, given new data
- the *prior* distribution represents our current beliefs, $P(Y)$
- the *posterior* distribution represents our beliefs after seeing the data
- the *likelihood* ($P(X|Y)$) is how likely we are to observe our new data, given our beliefs
- the *normalising constant* is $P(X)$; notice, independently of any Y, $P(X)$ will always be the denominator, which is why it is considered a "constant"

- **When are 2 RVs *marginally* independent?**
  - whenever an event happening doesn't affect the probability of another event happening. In other words, X and Y are *marginally independent* iff:
    $$P(X|Y) = P(X)$$
  - notice that this happens iff:
    $$P(X, Y) = P(X)P(Y)$$

- **When are 2 RVs *conditionally* independent?**
  - whenever knowing Y is sufficient to understand what happens to X, independently of some other observation Z:
    $$P(X|Y, Z) = P(X|Y)$$
    here we say "X is conditionally independent of Z given Y"
  - conditional independence does **not** imply marginal independence an viceversa
  - ***Example***: let S be the event that it is sunny, let B be the event that you go to the beach, and let H be the event that you get a heatstroke. Then:
    - it is likely that B and H are **not** marginally independent:
      $$P(B, H) > P(B)P(H)$$
      since it is more likely that you get a heatstroke as a consequence of going to the beach, than the heatstroke being caused by some other external factor, at the same time as going to the beach
    - if we know it is sunny, then it is perfectly likely that, as a consequence, we go to the beach **and** get the heatstroke, so knowing S is sufficient to explain H, without needed B:
      $$P(B, H|S) = P(B|H)P(B|S)$$

1.5 Gaussian Distribution

- **What is the one-dimensional Gaussian Distribution?**
  - if a random variable follows a *Gaussian/Normal Distribution* with mean μ and variance σ^2, then we say:

$$X \sim N(\mu, \sigma^2)$$

  - the PDF of a Normal Distribution is:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

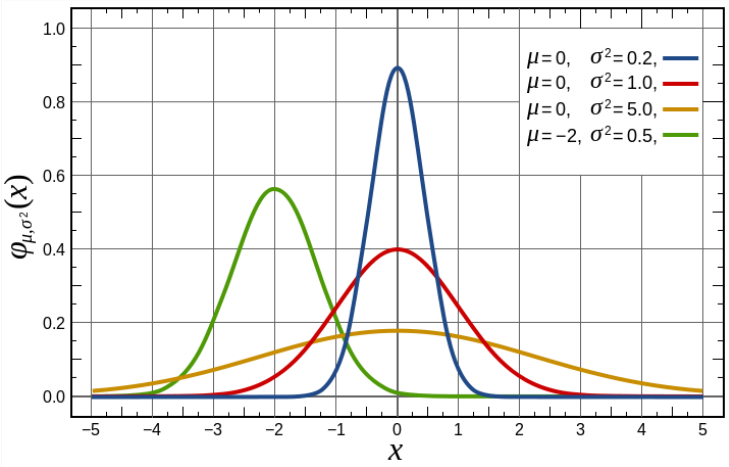
  - a standard normal distribution is one such that:

$$Z \sim N(0, 1)$$

  - any normal RV can be standardised via:

$$Z = \frac{X - \mu}{\sigma}$$

  - μ defines the symmetry line of the normal distribution
  - σ defines the width of the normal distribution



- **What is a covariance matrix?**
  - a useful structure when dealing with distributions in higher dimensions
  - to compute the covariance matrix. if we have 2 RVs, X and Y:

$$\Sigma = E\left[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T\right]$$

– indeed, if we compute this, we get:

$$\Sigma = \begin{pmatrix} E\left[(x_1-\mu_1)^2\right] & E\left[(x_1-\mu_1)(x_2-\mu_2)\right] \\ E\left[(x_1-\mu_1)(x_2-\mu_2)\right] & E\left[(x_2-\mu_2)^2\right] \end{pmatrix} = \begin{pmatrix} Var(x_1) & Cov(x_1,x_2) \\ Cov(x_1,x_2) & Var(x_2) \end{pmatrix}$$

- **What is the two-dimensional Gaussian Distribution?**
  - if $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, and they are independent, then the two-dimensional Gaussian is determined by the product of the PDFs:

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right)$$

  - this can also be expressed by using vectors:

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

    where Σ is the *covariance matrix*
  - using the vectorised form above:

$$\underline{x} - \underline{\mu} = \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}$$

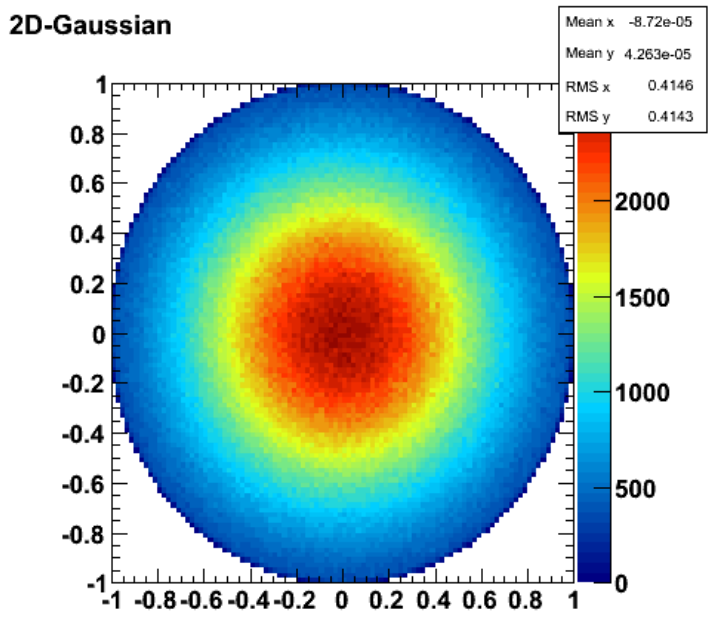
$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}$$

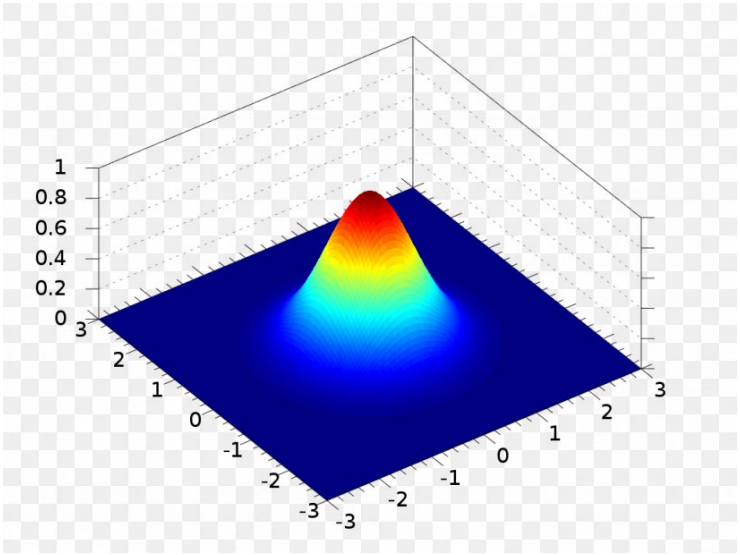
$$\Sigma^{-1}(\underline{x} - \underline{\mu}) = \begin{pmatrix} \frac{x_1-\mu_1}{\sigma_1^2} \\ \frac{x_2-\mu_2}{\sigma_2^2} \end{pmatrix}$$

$$(\underline{x} - \underline{\mu})^T\Sigma^{-1}(\underline{x} - \underline{\mu}) = \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}$$

    Thus:

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})^T\Sigma^{-1}(\underline{x} - \underline{\mu})\right)$$





- **What is the general, multivariate Gaussian Distribution?**
  - the multivariate Gaussian can be easily defined using the vector no-

tation above, with $\underline{x} \in \mathbb{R}^n$:

$$p(\underline{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\prod_i^n \sigma_i^2}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})\right)$$

- here $\underline{x} \sim N(\underline{\mu}, \Sigma)$

1.6 Estimating a Distribution

- **How can we derive a distribution from data (as opposed to using a distribution to produce data)?**
  - we aim at *learning* a distribution which could produce the data that we observe
  - the "best" distribution is that which is more probable to produce the observed data
- **How can we use a Maximum Likelihood Approach to estimate a distribution?**
  - asking "is this distribution more likely to produce this data?" is equivalent to maximising the probability density of observing the data (D), given a distribution (M):

$$P(D|M) = \prod_{i=1}^{N} P(D = \underline{x}_i|M)$$

  - for the above, we assume that each data point, $\underline{x}_i$ is independently generated, so we can simply multiply the probability of observing each point, given the distribution
  - this is the *Maximum Likelihood Approach*: selecting different models, and seeing which one is more likely to have produced the data
  - ***Example 1***: say we have the following data:

    10010101000001011101

    Since there are 2 outcomes, we consider a Bernoulli Distribution. To sample from the distribution, we consider 3 models:

    1. $M = 1$ (coin toss; H = 1, T = 1)
    2. $M = 2$ (die throw; 1 = 1, 2,3,4,5,6 = 0)
    3. $M = 3$ (double headed coin toss, H =1, T = 0)

    Letting c be the number of 1s, and using the Maximum Likelihood Approach:

$$P(D|M) = \prod_{i=1}^{N} P(D = \underline{x}_i|M) = P(D = 1|M)^c \times P(D = 0|M)^{20-c}$$

    Then:

1. if $M = 1$, $P(D = 1|M = 1) = 0.5$ and $P(D = 0|M = 1) = 0.5$, so the likelihood is:

$$0.5^{20} \approx 9.5 \times 10^{-7}$$

2. if $M = 2$, $P(D = 1|M = 2) = \frac{1}{6}$ and $P(D = 0|M = 1) = \frac{5}{6}$, so the likelihood is:

$$\frac{5^{11}}{6^{20}} \approx 1.3 \times 10^{-8}$$

3. if $M = 3$, $P(D = 1|M = 3) = 1$ and $P(D = 0|M = 3) = 0$, so the likelihood is 0

Thus, the most likely distribution is a fair coin toss. This can be generalised. Say the optimal distribution is such that:

$$P(D = 1|M) = \theta$$

Then,

$$P(D|M) = \theta^c \times (1 - \theta)^{n-c}$$

Taking the log, and calling the result $f(\theta)$:

$$f(\theta) = c \ln(\theta) + (n - c) \ln(1 - \theta)$$

The max likelihood will be the value of θ, such that $f'(\theta) = 0$, so:

$$f'(\theta) = \frac{c}{\theta} - \frac{n - c}{1 - \theta}$$

which means that:

$$\begin{aligned} & \frac{c}{\theta} - \frac{n - c}{1 - \theta} = 0 \\ \implies \quad & \frac{c}{\theta} = \frac{n - c}{1 - \theta} \\ \implies \quad & c - c\theta = n\theta - c\theta \\ \implies \quad & \theta = \frac{c}{n} \end{aligned}$$

which is an expected result

– ***Example 2***: we can also consider a model given by a Gaussian Distribution. If we have data with points x_i, then, lets assume that the model is generated with mean μ and variance σ^2. We seek to find the values of these parameters. To do this, lets consider the log

probability:

$$\begin{aligned}
\ln\left[P(D|M)\right] &= \ln\left[\prod_{i=1}^{N} P(D=x_i|\mu,\sigma^2)\right] \\
&= \sum_{i=1}^{N} \ln\left[P(D=x_i|\mu,\sigma^2)\right] \\
&= \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}})\right] \\
&= \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right] - \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{N} -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2
\end{aligned}$$

In order to get the max likelihood, we apply partial differentiation, as to get μ and σ^2.

$$\begin{aligned}
\frac{\partial}{\partial\mu}\left(\ln\left[P(D|M)\right]\right) &= \frac{\partial}{\partial\mu}\left(-\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right) \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu)
\end{aligned}$$

Setting this equal to 0:

$$\begin{aligned}
& \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0 \\
\implies \quad & \sum_{i=1}^{N}(x_i-\mu) = 0 \\
\implies \quad & \sum_{i=1}^{N} x_i = \sum_{i=1}^{N}\mu \\
\implies \quad & \mu = \frac{\sum_{i=1}^{N} x_i}{N}
\end{aligned}$$

We proceed similarly with variance:

$$\begin{aligned}\frac{\partial}{\partial\sigma}\left(\ln\left[P(D|M)\right]\right) &= \frac{\partial}{\partial\sigma}\left(-\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right)\\ &= -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2\\ &= \frac{-N\sigma^2 + \sum_{i=1}^{N}(x_i-\mu)^2}{\sigma^3}\end{aligned}$$

Setting this equal to 0:

$$\begin{aligned}&\frac{-N\sigma^2 + \sum_{i=1}^{N}(x_i-\mu)^2}{\sigma^3} = 0\\ \implies \quad &\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}\end{aligned}$$

All of the above generalises to multivariate Gaussians:

$$\underline{\mu} = \frac{\sum_{i=1}^{N}\underline{x}_i}{N}$$

$$\Sigma = \frac{\sum_{i=1}^{N}(\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})^T)}{N}$$

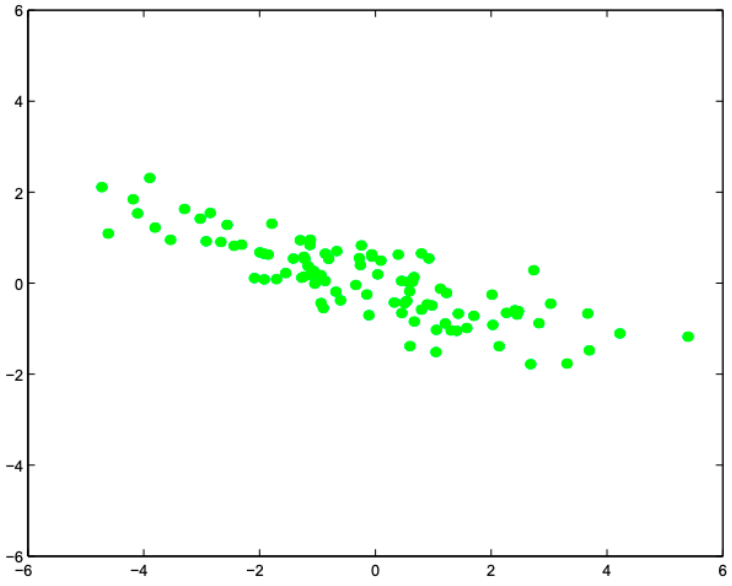


Figure 1: Data

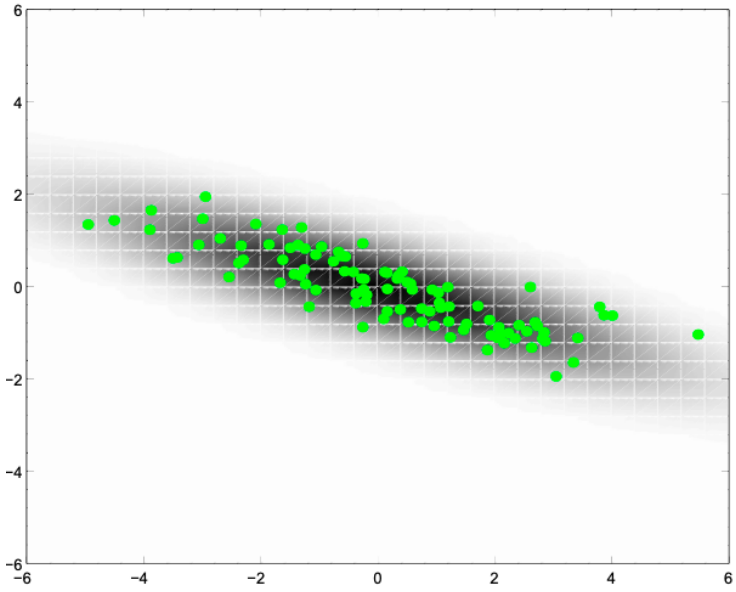


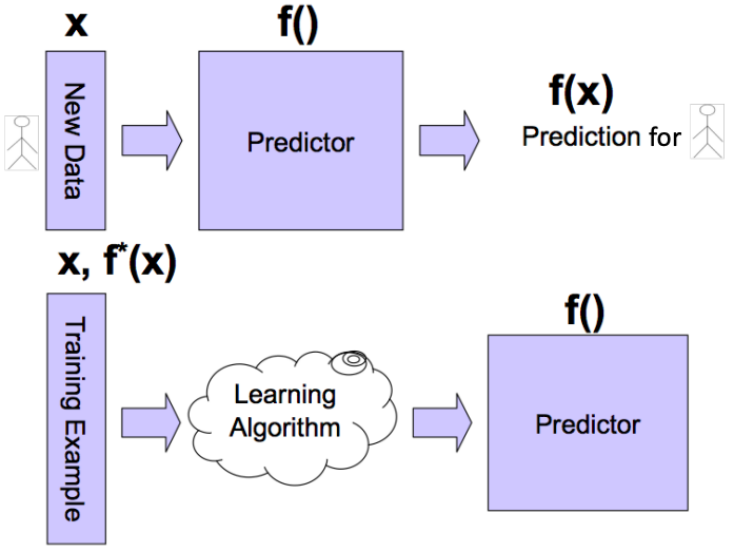
Figure 2: Bivariate Gaussian fitted over data

2 Thinking About Data

- The main tasks of ML are classification, regression and clustering
- Attribute-value pairs allow us to encode features as numerical values, which can be used by ML algorithms
- Models can be generative or discriminative, based on how the decision boundary is constructed

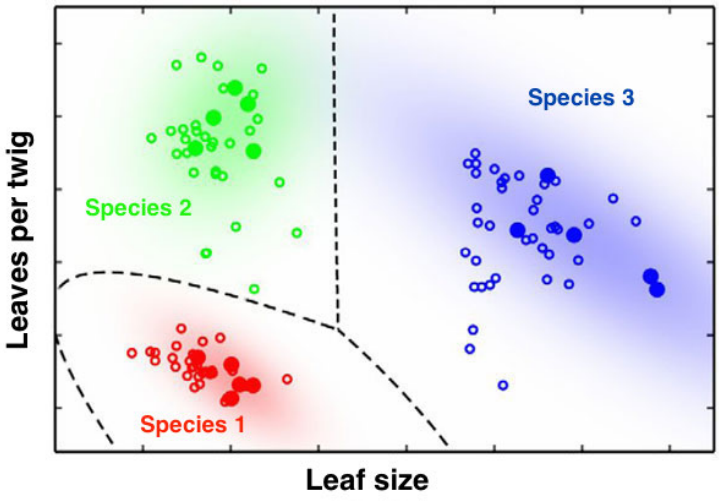
2.1 Aims of Machine Learning

- **What are learning algorithms in ML?**
  - ML seeks to learn from data, in order to make predictions
  - to make predictions, a *learning algorithm* is used to train a *predictor*, which is just a function

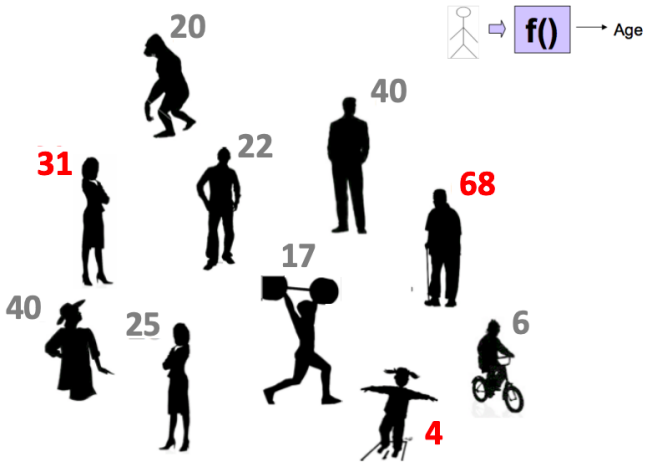


- **What are the 3 main tasks of ML?**

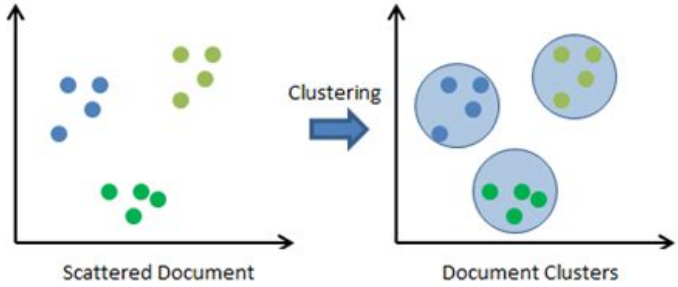
  1. **Classification**: labelling data with a certain class. Usually done by training and learning a *decision boundary* to split different classes up (*i.e determining whether an animal is a mammal, reptile or amphibian*)



  2. **Regression**: assigning numerical values to new data (*i.e predicting age given other characteristics, such as gender and pictures*)



3. **Clustering**: discover underlying patterns which allow you to split data into different clusters (*i.e discovering structure in the disposition of starts in the sky*)



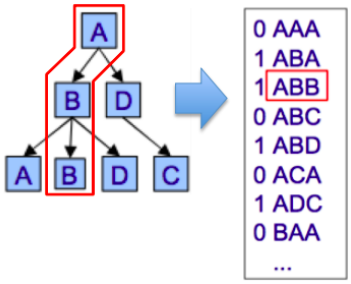
– classification and regression are *supervised* (require prior data labels), whilst clustering is *unsupervised*

2.2 Data as Attribute-Value Pairs

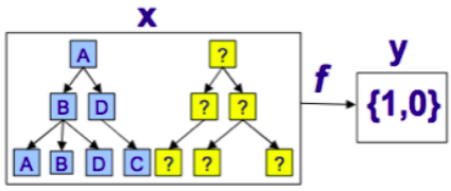
- **Why do we represent data as attribute-value pairs?**
  - since a predictor is a mathematical function, we need to be able to describe the object of our analysis (*i.e a human*) in more mathematical terms
  - objects can be described as an *unordered* bag of features (*i.e eye colour, height, occupation, etc ...*)
  - we consider 3 types of attributes:
    1. **Categorical**

2. **Ordinal**
3. **Numeric**

- **How can we embed structure within attribute-value pairs?**
  - attribute-value pairs are in essence unordered, but sometimes structure is necessary
  - embedding structure into the input can be done by treating attribute-value pairs as a tree, with the attributes being given as a root-to-leaf path:



  - embedding structure into the output can be done by building a predictor which takes an input tree and an output tree as inputs, and determines whether the provided output corresponds with the provided input. This now requires being able to search for plausible outputs.



- **What are the properties of categorical attributes?**
  - discrete
  - finite
  - unordered
  - mutually exclusive (synonyms can be a problem)
  - categories can be encoded as numbers (without meaning, only need to be able to apply $=, \neq$)
  - ***Example***: classical, jazz, rock, techno

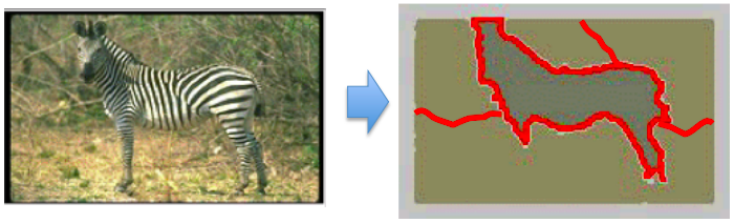
- **How do ordinal attributes differ from categorical attributes?**
  - ordinal categories have a natural ordering
  - numerical encodings must preserve this natural ordering (only to compare in size; no meaning to multiply/add/substract/...)
  - **Example**: primary, secondary, undergraduate, postgraduate
  - might be issues differentiating between categorical and ordinal (*i.e single, married, divorced*)
- **What are numerical attributes?**
  - real numbers
  - all operations have meaning (summation, multiplication, mean, variance)
  - numerical data is typically normalised, to ensure that scale does not affect the learning process
    - $x \in [0, 1]$
    - $\mu = 0, \sigma = 1$
- **What are the main issues related to using numerical attributes?**
  1. **Unusually Large/Small Values**: these mess up normalisation (all "normal" values will end up squashed up to one side of the scale upon normalisation), so must handle beforehand
  2. **Skewed Distributions**: caused by systematic, unusual values (*i.e personal wealth skews the distribution of wealth, but it can't be treated as an outlier, since it is a feature of the data*). Deal by applying $\log(x), \arctan(x)$ and then normalise. Otherwise it can affect regression, KNN, Naive Bayes.
  3. **Non-Monotonic Attributes**:
     - *monotone attribute*: direct correlation between attribute and value (*i.e higher net worth means lower lending risk*)
     - *non-monotone attribute*: correlation between attribute and value is not direct (*i.e age vs change of winning marathon. Being extremely young or old decreases chances of winning, need middle age*)
     - *quantisation* fixes non-monotonicity. Split values into categories (*i.e instead of considering age, use* < 20, $20 \leq x < 50$, ≥ 50)
     - affects regression, Naive Bayes

2.3 Picking Attributes

- **How should attributes be picked?**

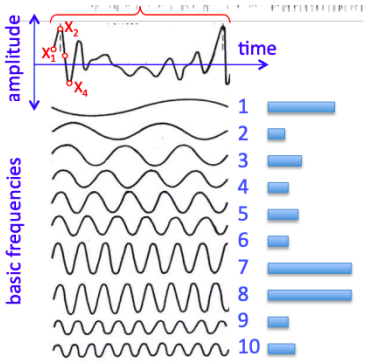
– sometimes straightforward: for credit scoring, we want to know about net worth, properties, job, age, etc...
– we should aim that similar objects have similar values associated with their attributes
– in other cases, the straightforward idea is not always the best:

1. **Digit Recognition**
   * can use pixels as attributes
   * pixels can be normalised
   * blurring image might improve performance (similar numbers might blur similarly)
   * this only works because we expect that numbers will have coloured pixels in the same positions
2. **Object Recognition**
   * pixels no longer work: a rotated zebra is still a zebra, so we need to handle rotations, translations, lighting, obstructions, etc ...
   * can use algorithms to segment image into regions, and then use the properties of the regions (i.e perimeter, area, colour frequency) to define attributes
   * if errors occur during segmentation, hope they are systematic



3. **Text Classification**
   * naively may use each word in the text as a value
   * main issue is that removing any word would completley shift the values that a machine receives (*i.e ["hello", "how", "are", "you"] and ["hello", "are", "you"] would be perceived as extremely different*)
   * instead, create a vocabulary of all the words, and return a binary string, with 1 if a word appears, and 0 if it doesn't
4. **Music Classification**
   * naively may use each point in the sound wave
   * main issue is that properties like amplitude can change, without necessarily changing the sound wave. Waves can be shifted, and they would be perceived as different.

* instead, use Fourier Transform, and use the weights as the attributes



2.4 Supervised vs Unsupervised Learning

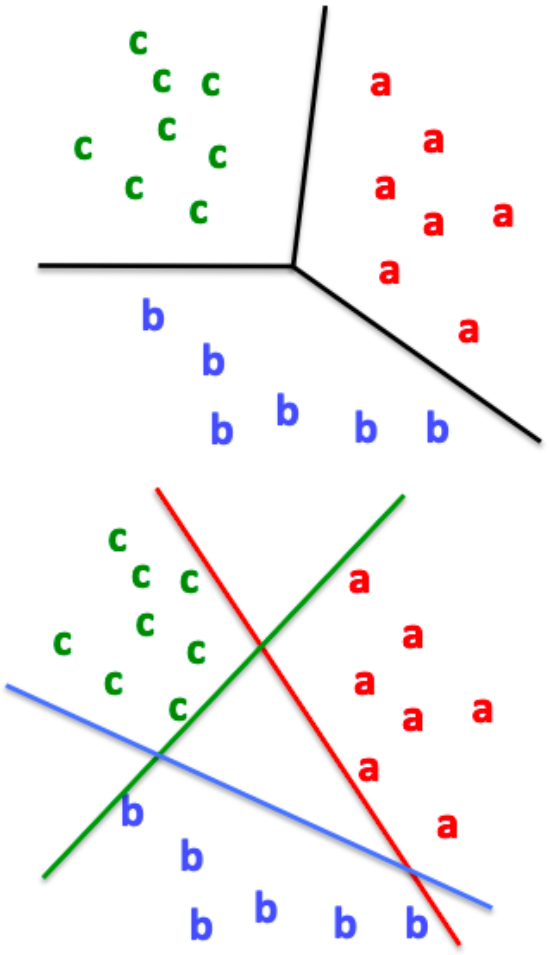
- **What defines a supervised learning algorithm in ML?**
  - training data has labels
  - use these labels to predict for new data
  - can measure accuracy directly
- **What are the differences between supervised and unsupervised learning algorithms?**
  - rather than labelling, seeks to understand the (underlying) structure of data
  - labelling not required during training
  - require indirect or qualitative (are these the results I wanted?) evaluation
- **How does semi-supervised learning compare to supervised and unsupervised learning?**
  - semi-supervised learning employs unsupervised learning to improve supervised learning
  - small number of labelled data + lots of unlabelled data

2.5 Multi-class vs Binary Classification

- **What is multi-class classification?**
  - produce decision boundary to separate many classes
  - classes are mutually exlusive and exhaustive

- for example, Naive Bayes, KNN, decision tree, logistic (softmax extension)

- **What is binary classification?**
  - one class vs the rest (class A vs class not A)
  - classes can overlap (a region might be included within 2 decision boundaries)
  - SVM, logistic (basic), perceptron



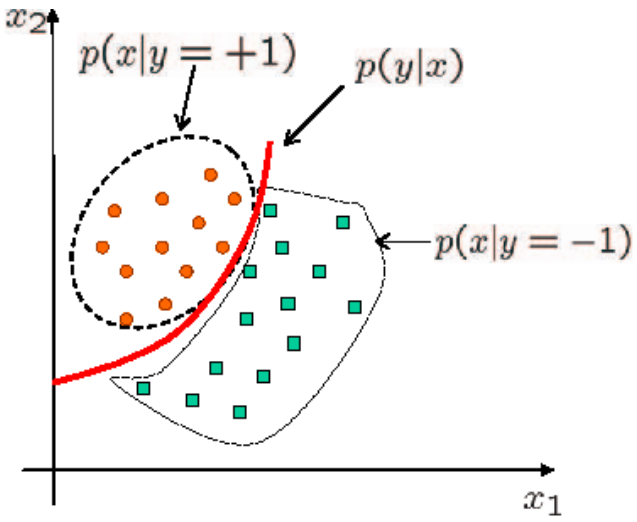
2.6 Accuracy and Unbalanced Classes

- **Is accuracy the best metric to measure ML success?**
  - a more accurate classifier may not always be a better classifier

- if classes are very unbalanced, then a classifier can always guess the same class, and have a high accuracy
- ***Example***: predicting if a paper will win a Nobel prize. It is so rare, that we can achieve more than 99% accuracy by simply labelling every paper as "non-Nobel prize"
- might be smoothed by making the cost of a false negative bigger than the cost of a false positive

2.7 Generative vs Discriminative

- **What is a *generative* ML model?**
  - generates a region for a class
  - in essence, constructs a probabilistic model to describe in what region a class is most likely to appear
  - lends itself to using unlabelled data
- **What is a *discriminative* ML model**
  - focuses on deriving a decision boundary for classes
  - requires many labelled examples



ustration of generative v.s. discriminative models. Discri

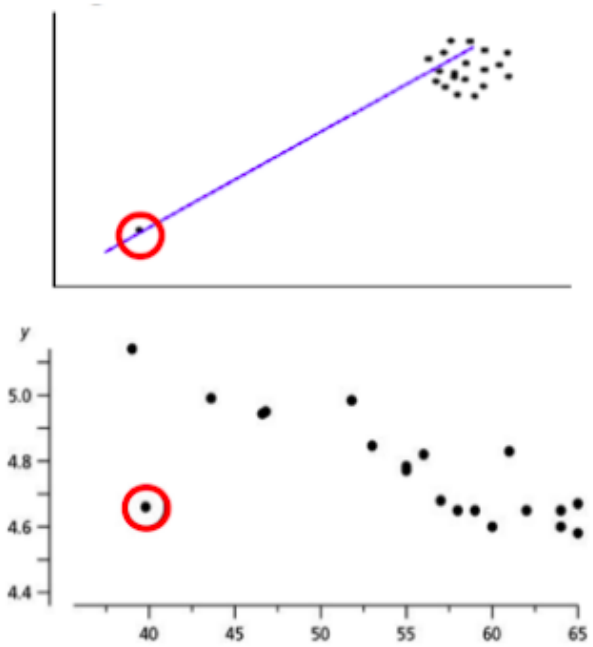
2.8 Data Outliers

- **What are outliers?**
  - isolated instance of a class, different from all other instance

  - affects performance of ML models

- **How do we deal with extreme outliers?**
  - easy to detect (i.e via confidence intervals)
  - remove
  - enforce a threshold for all the data

- **How do we deal with "hidden" outliers?**
  - harder to detect
  - might need to visualise to check



3 Naive Bayes

- A Naive Bayes classifier is a Bayesian Classifier which assumes that data features are conditionally independent, given a class
- In Gaussian Naive Bayes, we assume that each feature is modelled by its own Gaussian Distribution
- Naive Bayes is hindered by data which is different but has the same distribution, the zero frequency problem and the assumption of conditional independence. Nonetheless, it can cope with missing data.

3.1 Bayesian Classifiers

- **What is a probabilistic classifier?**
  - classifier which aims to assign classes, based on which class y is most likely to have produced the observed result $\underline{x}$:

$$\hat{y} = \arg\max_y P(Y = y|\underline{X} = \underline{x})$$

    (Here $\underline{x}$ represents all the observations made, so $\underline{x} = x_1, x_2, \ldots, x_n$)

- **What are the components of a Bayesian Classifier?**
  - a *Bayesian Classifier* is a probabilistic classifier which uses Bayes' Theorem to compute the conditional probability:

$$P(Y = y|\underline{X} = \underline{x}) = \frac{P(\underline{X} = \underline{x}|Y = y)P(Y = y)}{P(\underline{X} = \underline{x})} = \frac{P(\underline{X} = \underline{x}|Y = y)P(Y = y)}{\sum_{y'} P(\underline{X} = \underline{x}|Y = y')P(Y = y')}$$

  - for classifiaction, the *normalisation term* is not necessary (it'll be constant $\forall y$). It is nonethless useful when comparing elements within a class (*i.e given $\underline{x}_1$ and $\underline{x}_2$, which of the 2 is most likely to be suffering from a disease; outliers will tend to have much lower probabilities*)

3.2 Naive Bayes

- **What assumption defines a Naive Bayes Model?**
  - Naive Bayes assumes that each of the $x_1, x_2, \ldots, x_n$ are **conditionally independent given y**
  - this allows us to create a simplified model, which is less computationally expensive (*i.e consider computing the probability of seeing a specific pixel being turned on or off across a single 20×20 image. This has 2^{400} possibilities. We would then need to go through*

all 20×20 *images, and consider the probability of one of these* 2^{400} *pixels being turned on or off for a specific class guess*)

- in order to calculate $P(Y = y|\underline{X} = \underline{x})$, we need to first calculate $P(\underline{X} = \underline{x}|Y = y)$. By the definition of conditional probability:

$$P(x_1, x_2, \ldots, x_n|y) = \frac{P(x_1, x_2, \ldots, x_n, y)}{P(y)}$$

Using the Chain Rule (1.4), we know that:

$$P(x_1, x_2, \ldots, x_n, y) = \prod_{i=1}^{n} P(x_i|x_{i+1}, \ldots, x_n, y)$$

But since we are assuming conditional independence, this reduces to:

$$P(y) \times \prod_{i=1}^{n} P(x_i|y)$$

So:

$$P(x_1, x_2, \ldots, x_n|y) = \prod_{i=1}^{n} P(x_i|y)$$

Overall, our Naive Bayes model becomes:

$$P(Y = y|\underline{X} = \underline{x}) = \frac{P(Y = y) \times \prod_{i=1}^{n} P(x_i|y)}{\sum_{y'} P(\underline{X} = \underline{x}|Y = y')P(Y = y')}$$

- conditional independence is justified by the fact that the class y must be the underlying cause for each of the attribute values, and further, a attribute value x_i should have nothing to do with the value of x_2 (*i.e drawing a 3 in a bitmap image, it is the "threeness" that puts the pixels into the disposition*)

- **Is Naive Bayes generative or discriminative?**
  - we are creating a probabilistic model (distribution) for each class, so Naive Bayes is *generative*
  - in particular, this allows us to create synthetic class instances

3.3 Naive Bayes: Continuous Example

We aim to use Naive Bayes to classify an individual (as adult or child), based on weight and height. Our data is given by the following:

- h_i: the height of the individual [numerical]
- w_i: the weight of the individual [numerical]

- y_i: a (adult) or c (child) [categorical]

There are **4 adults** and **12 children**. Recall our Naive Bayes model:

$$P(Y = y|\underline{X} = \underline{x}) = \frac{P(Y = y) \times \prod_{i=1}^{n} P(x_i|y)}{\sum_{y'} P(\underline{X} = \underline{x}|Y = y')P(Y = y')}$$

Thus, we have the following components:

- $P(Y = a)$
- $P(Y = c)$
- $P(w_i|y_i)$
- $P(h_i|y_i)$
- $P(w_i, h_i|y_i) \times P(Y = y_i) = P(w_i|y_i) \times P(h_i|y_i) \times P(Y = y_i)$

The first 2 are straightforward:

$$P(a) = \frac{4}{4 + 12} = \frac{1}{4}$$
$$P(c) = \frac{12}{4 + 12} = \frac{3}{4}$$

For the remaining, we need to choose a model. Since we are dealing with weights and heights, it makes sense to consider a Gaussian Distribution for each. Thus, for adults:

$$\mu_{h,a} = \frac{1}{4} \times \sum_{i,y_i=a} h_i$$
$$\sigma^2_{h,a} = \frac{1}{4} \times \sum_{i,y_i=a} (h_i - \mu_{h,a})^2$$
$$\mu_{w,a} = \frac{1}{4} \times \sum_{i,y_i=a} w_i$$
$$\sigma^2_{w,a} = \frac{1}{4} \times \sum_{i,y_i=a} (w_i - \mu_{w,a})^2$$

and for children:

$$\mu_{h,c} = \frac{1}{12} \times \sum_{i,y_i=c} h_i$$
$$\sigma^2_{h,c} = \frac{1}{12} \times \sum_{i,y_i=c} (h_i - \mu_{h,c})^2$$

$$\mu_{w,c} = \frac{1}{12} \times \sum_{i,y_i=c} w_i$$

$$\sigma^2_{w,c} = \frac{1}{12} \times \sum_{i,y_i=c} (w_i - \mu_{w,c})^2$$

Then the properties for our model can be easily computed. Define:

$$Gaussian(x, \mu, \sigma^2)$$

as the Gaussian for the parameters. Then:

$$P(w_i|a) = Gaussian(w_i, \mu_{w,a}, \sigma^2 w, a)$$
$$P(h_i|a) = Gaussian(h_i, \mu_{h,a}, \sigma^2 h, a)$$
$$P(w_i|c) = Gaussian(w_i, \mu_{w,c}, \sigma^2 w, c)$$
$$P(h_i|c) = Gaussian(h_i, \mu_{h,c}, \sigma^2 h, c)$$
$$P(w_i, h_i|a) = P(w_i|a) \times P(h_i|a) \times P(a)$$
$$P(w_i, h_i|c) = P(w_i|c) \times P(h_i|c) \times P(c)$$

So for example:

$$P(a|w_i, h_i) = \frac{P(a) \times P(w_i|a) \times P(h_i|a)}{P(w_i|a) \times P(h_i|a) \times P(a) + P(w_i|c) \times P(h_i|c) \times P(c)}$$

The above model can result in the following:

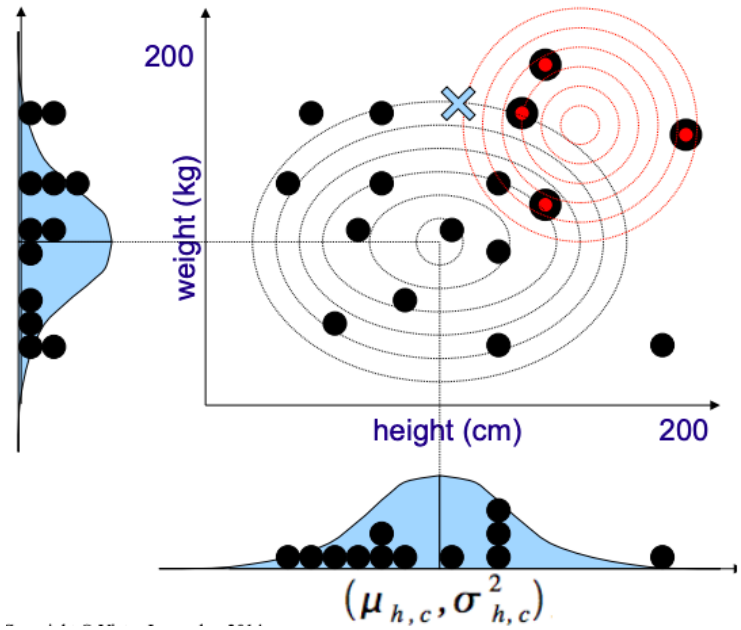


Figure 3: If we want to classify the X, notice its height is normal for a child, but the individual is overweight. ON the other hand, it has the right weight for an adult, but it is not as tall as expected. We would need to calculate probabilities in order to make a sensible decision.

3.4 Naive Bayes: Discrete Example

Naive Bayes can also be employed in discrete examples. For example, consider a spam classifier. The following is our data:

send us your password	**spam**
send us your review	ham
review your password	ham
review us	**spam**
send your password	**spam**
send us your account	**spam**

Using attributes as words, we can obtain the following table of probabilities:

word	$P(word\|spam)$	$P(word\|ham)$
-	4/6	2/6
password	2/4	1/2
review	1/4	2/2
send	3/4	1/2
us	3/4	1/2
your	3/4	1/2
account	1/4	0/2

Say we get the message:

review us now

Firstly, since "now" is not part of our vocabulary, we just ignore it (see further on whether this affects results).

Thus, we need only consider the phrase "review us". Whenever we see "review" or "us", we use $P(word|spam/ham)$, but for all the other words in the vocabulary, we consider $1 - P(word|spam/ham)$:

$$P(review\ us|spam) = \left(1 - \frac{2}{4}\right)\left(\frac{1}{4}\right)\left(1 - \frac{3}{4}\right)\left(\frac{3}{4}\right)\left(1 - \frac{3}{4}\right)\left(1 - \frac{1}{4}\right) = 0.0044$$

$$P(review\ us|ham) = \left(1 - \frac{1}{2}\right)\left(\frac{2}{2}\right)\left(1 - \frac{1}{4}\right)\left(\frac{1}{2}\right)\left(1 - \frac{1}{2}\right)\left(1 - \frac{0}{2}\right) = 0.0625$$

Hence, the probability that "review us now" is ham, given all the above data is:

$$\begin{aligned} P(ham|review\ us\ now) &= \frac{P(review\ us|ham) \times P(ham)}{P(review\ us|ham) \times P(ham) + P(review\ us|spam) \times P(spam)} \\ &= \frac{0.0625 \times \frac{2}{6}}{0.0625 \times \frac{2}{6} + 0.0044 \times \frac{4}{6}} \\ &= 0.87 \end{aligned}$$

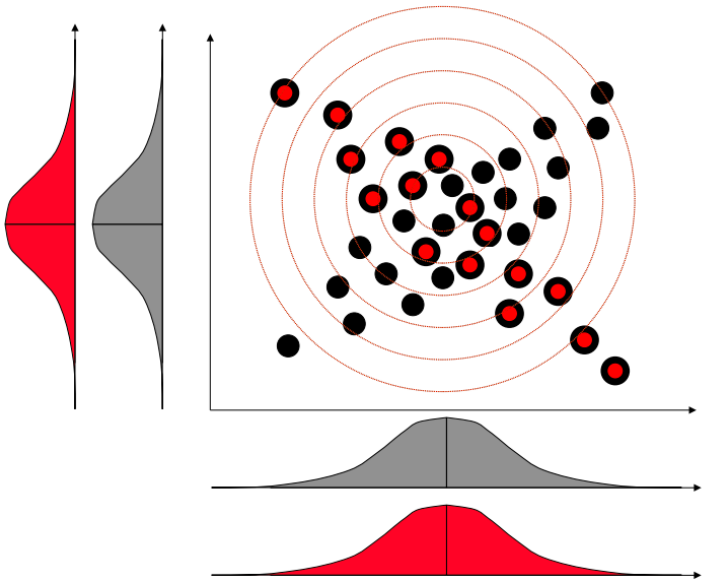
But notice: the message "review us" was spam in the data, but NB labels it as ham. This is a consequence of the independence assumption.

3.5 Issues With Naive Bayes

- **What 3 issues affect Naive Bayes?**

  1. **Equal Distributions, Different Data**
     - it is well known that different data can have identical summary statistics

- thus, NB won't work on data that is inherently different, but have the same mean and variance



2. **Zero-Frequency Problem**
   - it is possible that we encounter new words that are not in our vocabulary (by Zipf's Law, 50% of words will be new)
   - this can lead to issues, particularly that some probabilities will be set to 0 (for example, "account" did not appear in any ham sentence)
   - apply *Laplace smoothing* (add small quantity to all counts)
3. **Conditional Independence**
   - in some cases, this is not a valid assumption
   - for example, in sentences, word often times go together
   - we can fool a NB classifier by including lots of instances of one class
   - ***Example***: in the above discrete case, "review us" was flagged as ham because "review" appears in all of the ham instances

3.6 Naive Bayes and Missing Data

- **Is Naive Bayes reliable, even if data is missing?**
  - Yes! If data is missing in Naive Bayes, the strategy is to simply ignore it in the computation (as we did with "now" in the discrete example)

– say you have data, but the value of X_j is missing:

$$X_1 = x_1, \ldots, X_j = ??, \ldots X_n = x_n$$

But then the probability of the event is going to be:

$$\sum_{x_j} P(x_1, \ldots, x_j, \ldots, x_n)$$

– if we apply the idea above to Naive Bayes:

$$\begin{aligned} P(x_1, \ldots, x_j = ?, \ldots, x_n|y) &= \sum_{x_j} \left(\prod_{i=1}^{n} P(x_i|y) \right) \\ &= P(x_1|y) \times \ldots \times \left(\sum_{x_j} P(x_j|y) \right) \times \ldots \times P(x_n|y) \\ &= P(x_1|y) \times \ldots \times 1 \times \ldots \times P(x_n|y) \\ &= \prod_{i=1, i \neq j}^{n} P(x_i|y) \end{aligned}$$

which is the same as if we had completely ignored x_j