AINT351 - Revision for Ian

Three types of learning

Imagine a machine experiences a sequence of sensory inputs `x1, x2, ... xn`

Supervised learning:

- The machine is also given `y1, y2, ..yn` and its goal is to learn and reproduce them from the inputs
- Learning by examples, input and output is given so it knows how to reproduce the output from the input

Unsupervised learning:

- The machine should build a representation of x that can be used for decision making, prediction
- There is no desired output, you are given inputs and after some iterations you start to categorise data based on some criteria

Reinforcement learning:

- The machine can generate actions `a1, a2,.. an` that affects its environment and receives a reward or punishment based on them. Its goal is to learn actions that maximise long term reward
- Learning based on rewards for actions so that it learns to maximise long term reward

Goals of supervised learning

Classify input data:

- In this case the desired outputs `y1,y2,...yn` are discrete class labels and the goal is to **classify** new output correctly from the new input
- have an image of a digit and want to know what digit it is based on previous examples of that digit

Goals of unsupervised learning

Regression

- In this case the desired outputs `y1, y2, ...yn` are continuous values and the goal is to **predict** new output correctly from new input
- Have the data from babies and can try to predict its weight given its height

We wish to find useful representations of data. This can involve

- Finding clusters
- Dimensionality reduction

- Finding the hidden cause of the surface phenomena
- Modelling the data probability density
- Data compression

Probability

Types of data:

Discrete data: only certain values

- Dice value = {1,2,3,4,5,6}
- Flip a coin = {H,T}

Continuous data: any value

- Length measurement
- Weight measurement

Probability functions

- A probability function maps possible values of a variable to its respective probabilities
  - e.g. if value is x we can write its possible probabilities as p(x)
- Probability functions have the following properties
  - P(x) is a number with a value between 0 to 1.0
  - The area under a probability function is always unity

The addition law of probability

- If two events A and B are mutually exclusive then
- `P(A U B)` = the probability event A **OR** B occurs
- `P(A U B) = P(A) + P(B)`
- If two events A and B are **NOT** mutually exclusive then
- `P(A U B) = P(A) + P(B) - P(A n B)`
- You have to subtract the intersect as it is where both events happen

Probability distributions

Bernoulli distribution:

- The probability of a success of failure, heads or tails, 1 or 0
- `n` is the number of times that the experiment is repeated

Discrete distribution:

- A finite amount of probabilities all of which have equal probability of occurring
- A dice throw, each outcome has a probability = 1/6

Cumulative probability:

- The probability of this event happening **ASWELL AS** all the previous events
- A dice landing on 6 as well as all the chances of it landing on 1,2,3,4 and 5 = 6/6

Binomial distribution:

- 2 outcomes
  - Heads or tails
- What is the probability of getting exactly 3 heads in 5 coin tosses
- HHHTT
  - `(1/2)^3 x (1/2)^2`
- THHHT
  - `(1/2)^1 x (1/2)^3 x (1/2)^1`
- All equal = `(1/2)^3 x (1/2)^2`
- therefore the overall probability =
  - `N x (1/2)^3 x (1/2)^2`
  - where N = number of unique arrangements
- There are exactly 10 ways to get 3 heads in 5 coin tosses
  - N = 10
- `10 x (1/2)^3 x (1/2)^2 = 0.3125`

Uniform distribution:

- A distribution that has constant probability
- 0.5 of values 0 and 1.0

Continuous data distributions:

- A continuous random variable is a random variable with a set of possible values that is infinite or uncountable
- looks like Gaussian distribution

Variance

- It measures how far a set of random numbers are spread out from their mean
- variance is the expectation of the squared deviation of a random variable from its mean

Expected value

- What is the expected value of `x` given it is in a certain area under a curve
- To find:
  - Get the marginal distribution by summing all the probabilities in that row or column (for that value of x or y)
  - Multiply the value of x by the marginal distribution of that value

Covariance: Joint probability

- The covariance measures the strength of the linear relationship between two variables
- to get the covariance of expected values:
  - find the expected value of x and expected value of y by using steps described above
  - get the expected value of xy = E(XY) by multiplying each value in the table by its x and y values
  - and then plug all values into equation
  - `COV(XY) = E(XY) - E(X) . E(Y)`

Coefficient of XY

- To get the coffecient of XY we need the standard devation of x and standard deviation of y as well as the covariance figured out in steps above
- 'Coff(XY) = COV(XY) / STD(X) . STD(Y)
- First get the variance of x and y
  - `variance of x = E(X^2) - E(X)^2` - to get the `E(X^2)` need to do it the same as the expected values by square all the values of x
  - `variance of y = E(Y^2) - E(Y)^2` - to get the `E(Y^2)` need to do it the same as the expected values by square all the values of Y
- Once you have the variance you can square root it to get the standard deviation
- then plug it all back into the equation

Conditional probability distribution, independence

- How can you tell if x and y are independant
- Changing the value of y should have no effect on the probability distribution of x

Effect of standard deviation

- The variance either side of the mean

- The greater the standard deviation the greater the probability that x is within it
- Greater the standard deviation the further it is from the mean

Cumulative distribution function

- For a continuous random variable X the cumulative distribution function
- CDF(x) represents the area under the probability density function P(x) to the left of X
- `CFD(x) = P(X < x)`

Exponential distribution

- Given `P(x) = e^-x`
- What is the probability of x falling withing 1 to 2
- `P(1 <= x <= 2)` corresponds to area under distribution between 1 and 2

Marginalization

- sum up all the columns and row values for either x or y respectively

Conditional probability: bayes rule

- 'P(A n B) is the probability of A and B happening
- can be written P(A, B) - the joint distribution of A and B
- The probability of A happening multiplied by the probability of B happening given that A has happened

$$P(A \cap B) = P(B|A)P(A)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- if A and B are **NOT independent events** then - P(A n B) = P(A) . P(A|B)
- **This relationship is always true**

Interpreting covariance

- If we calculate the covariance between two random variables
- if the cov(X, Y) > 0
  - X and Y are positively correlated
- If the cov(X, Y) < 0
  - X and Y are inversely correlated
- If the cov(X, Y) $= 0$
  - X and Y are independent

Marginalisation

- Sum of probabilities across their given variable

Conditional probability

- If A and B are **NOT** independent events then
- `P(A n B) = P(A) . P(B|A)`
- This relationship is always true
- A has to happen for it to be true
- If A and B **ARE** independent events then
- P(A n B) = P(A) . P(B)
- This relationship is only true if A and B are independent
- Rolling a dice:
  - The first throw doesn't effect the second

Product rule of probability

- Product rule states that
- `P(A,B) = P(B)P(A|B)`
- `P(A,B) = P(A)P(B|A)`
  - The joint probability of A and B is prob of A multiplied by the probability of A given B
    - Same the other way around
- Leads to bayes rule
- `P(B)P(A|B) = P(A)P(B|A)`
- `P(A|B) = P(A)P(B|A)/P(B)`
  - The probability of A given B is equal to:
  - The probability of A multiplied by the probability of B given A

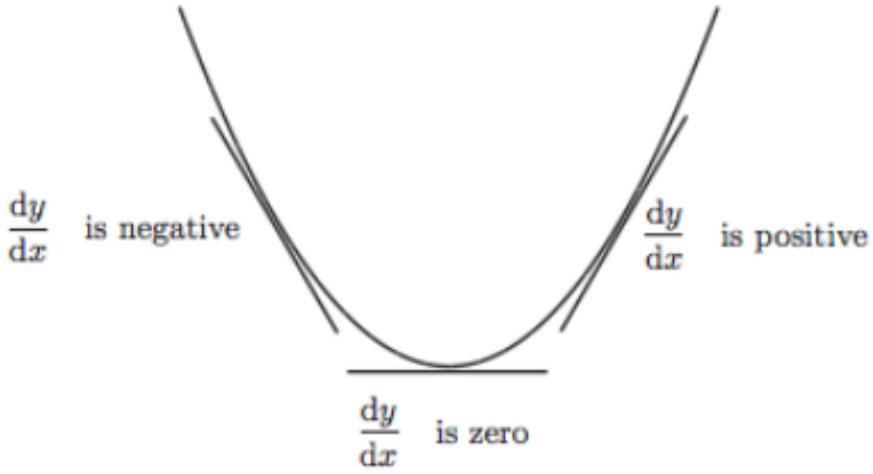
  - Over the probability of B
  - This is bayes theorem
- If A and B are independent
  - `P(A, B) = P(A)P(B)`
  - Joint probability of two independent events is the dot product probability of both events happening

Gradient descent

- How can we get to the top of a hill?
  - We follow the gradient
- How can we get to the bottom of a hill
  - We follow the descent
- How do we know we're at the top
  - It goes down on both sides

Gradient of a curve

- Gradient is the slope of a curve or surface
- Going up the hill it is +ve
- Going downhill it is -ve
- Differentiation finds tagent of line on graph
- Gradient of straight line =
  - Change in y
  - over change in x



Iterative gradient decent

- Find minimum of a function

- Move downwards in direction of gradient

Local maxima and minima

- Can be local minima as well as global minima
  - Same for maxima
- Gradient descent can get stuck in local minima/maxima

Least squares fitting

- We want to fit a straight line to data measurements
- Equation for straight line in a single dimension is:
  - yi = mxi + c
- Sum error over all points