IAML: Basic Maths, Probability and Estimation

Nigel Goddard
School of Informatics

Semester 1

Why Maths?

- IAML is focused on intuition and algorithms, not theory
- But sometimes you need maths to express the algorithms
- e.g., We represent training instances via vectors ($\mathbf{x} \in \mathbb{R}^k$), and linear functions of them as matrices
- Your first-year courses covered this stuff
  - But unlike many Informatics courses, we actually use it!

Functions, logarithms and exponentials

- Defining functions.
- Variable change in functions.
- Evaluation of functions.
- Combination rules for exponentials and logarithms.
- Properties of exponential and logarithm.

Vectors

- Scalar (dot) product, transpose.
- Basis vectors, unit vectors, vector length.
- Orthogonality, gradient vector, planes and hyper-planes.

Matrices

- Matrix addition, multiplication
- Matrix inverse, determinant.
- Linear transformation of vectors
- Eigenvalues, eigenvectors, symmetric matrices.

Calculus

- General rules for differentiation of standard functions, product rule, function of function rule.
- Partial differentiation
- Definition of integration
- Integration of standard functions.

Probability and Statistics

We will go over these, but useful if you have seen these before.

- Probability, events
- Mean, variance, covariance
- Conditional probability
- Combination rules for probabilities
- Independence, conditional independence

Why Probability?

Probability is a branch of mathematics concerned with the analysis of uncertain (random) events

Examples of uncertain events

- Gambling: Cards, dice, etc.
- Whether my first grandchild will be a boy or a girl[1]
- The number of children born in the UK last year
- The title of the next slide

Notice that

- Uncertainty depends on what you know already
- Whether something is "uncertain" is a pragmatic decision

[1] I have no grandchildren currently, but I do have children

Why Probability in Machine Learning?

The training data is a source of uncertainty.

- Noise. e.g., Sensor networks, robotics
- Sampling error. e.g., Choice of training documents from the Web

Many learning algorithms use probabilities explicitly

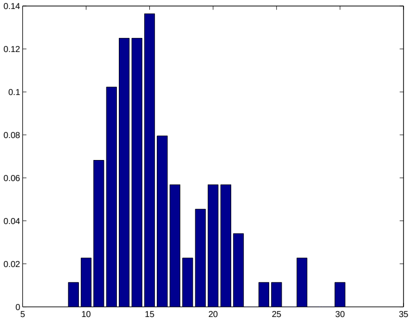
Ones that don’t are still often *analyzed* using probabilities.

Random Variables

- The set of all possible outcomes of an experiment is called the *sample space*, denoted by Ω
- Events are subsets of Ω (often singletons)
- A random variable takes on values from a collection of *mutually exclusive* and *collectively exhaustive* states, where each state corresponds to some event
- A random variable X is a map from the sample space to the set of states
- Examples of variables
  - Colour of a car *blue*, *green*, *red*
  - Number of children in a family $0, 1, 2, 3, 4, 5, 6, > 6$
  - Toss two coins, let $X = (\text{number of heads})^2$. What values can X take?

Discrete Random Variables

Random variables (RVs) can be *discrete* or *continuous*.



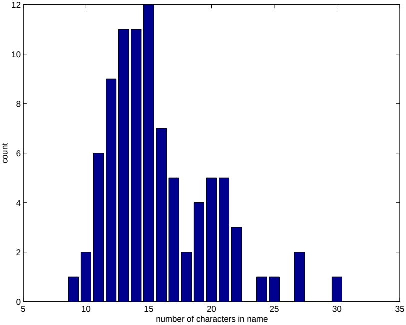
- Use capital letters to denote random variables and lower case letters to denote values that they take, e.g. $p(X = x)$. Often shortened to $p(x)$.
- $p(x)$ is called a *probability mass function*.
- For discrete RVs: $\sum_x p(x) = 1$.

Examples: Discrete Distributions

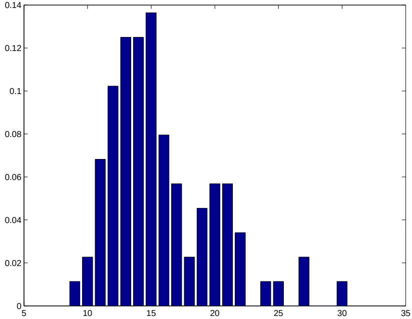
- Example 1: Coin toss: 0 or 1
- Example 2: Have data for the number of characters in names of 88 people submitting tutorial requests:

  `9 10 10 11 11 11 11 11 11 12 12 12 12 12 12`
  `12 12 12 13 13 13 13 13 13 13 13 13 13 13`
  `14 14 14 14 14 14 14 14 14 14 14 15 15 15`
  `15 15 15 15 15 15 15 15 15 16 16 16 16 16`
  `16 16 17 17 17 17 17 18 18 19 19 19 19 20`
  `20 20 20 20 21 21 21 21 21 22 22 22 24 25`
  `27 27 30`
- Example 3: Third word on this slide.

Frequency



frequency



normalized frequency

Joint distributions

- Suppose X and Y are two random variables. X takes on the value *yes* if the word "password" occurs in an email, and *no* if this word is not present. Y takes on the values of *ham* and *spam*
- This example relates to "spam filtering" for email

	$Y = ham$	$Y = spam$
$X = yes$	0.01	0.25
$X = no$	0.49	0.25

- Notation
  $p(X = yes, Y = ham) = 0.01$

Marginal Probabilities

The *sum rule*

$$p(X) = \sum_{y} p(X, Y)$$

e.g. $P(X = yes) = ?$

Marginal Probabilities

The *sum rule*

$$p(X) = \sum_{y} p(X, Y)$$

e.g. $P(X = yes) = ?$

Similarly:

$$p(Y) = \sum_{x} p(X, Y)$$

e.g. $P(Y = ham) = ?$

Conditional Probability

- Let $\mathbf{X}$ and $\mathbf{Y}$ be two disjoint subsets of variables, such that $p(\mathbf{Y} = \mathbf{y}) > 0$. Then the *conditional probability distribution* (CPD) of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ is given by

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

- Gives us the *product rule*

$$p(\mathbf{X}, \mathbf{Y}) = p(\mathbf{Y}) p(\mathbf{X}|\mathbf{Y}) = p(\mathbf{X}) p(\mathbf{Y}|\mathbf{X})$$

- **Example**: In the ham/spam example, what is $p(X = yes | Y = ham)$?
- $\sum_{\mathbf{x}} p(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}) = 1 \text{ for all } \mathbf{y}$

Bayes' Rule

- From the product rule,

$$p(\mathbf{Y}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})}{p(\mathbf{X})}$$

- From the sum rule the denominator is

$$p(\mathbf{X}) = \sum_{y} p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})$$

- Say that **Y** denotes a class label, and **X** an observation. Then $p(\mathbf{Y})$ is the *prior* distribution for a label, and $p(\mathbf{Y}|\mathbf{X})$ is the *posterior* distribution for **Y** given a datapoint **x**.

Independence

- Independence means that one variable does not affect another, X is *(marginally) independent* of Y if

$$p(X|Y) = P(X)$$

- This is equivalent to saying

$$p(X, Y) = p(X)p(Y)$$

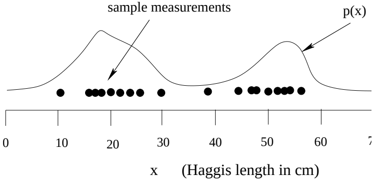
  (can show this from definition of conditional probability)
- X_1 is *conditionally independent* of X_2 given Y if

$$p(X_1|X_2, Y) = p(X_1|Y)$$

  (i.e., once I know Y, knowing X_2 does not provide additional information about X_1)
- These are different things. Conditional independence does not imply marginal independence, nor vice versa.

Continuous Random Variables

Suppose we want random values in $\mathbb{R}$. Example:



- Formally, a continuous random variable X is a map $X : \Sigma \to \mathbb{R}$.
- In continuous case, $p(x)$ is called a *density function*
- Get the probability $\Pr\{X \in [a, b]\}$ by integration

$$\Pr\{X \in [a, b]\} = \int_a^b p(x)dx$$

- Always true: $p(x) > 0$ for all x and $\int p(x)dx = 1$ (*cf* discrete case).
- Bayes' rule, conditional densities, joint densities work exactly as in the discrete case.

Mean, variance

For a continuous RV

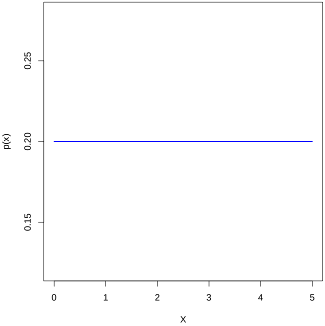
$$\mu = \int xp(x)dx \qquad \sigma^2 = \int (x-\mu)^2 p(x)dx$$

- μ is the *mean*
- σ^2 is the *variance*
- For numerical discrete variables, convert integrals to sums
- Also written: $EX = \int xp(x)dx$ for the mean and
- $VX = E(X-\mu)^2 = \int (x-\mu)^2 p(x)dx$ for the variance

Example: Uniform Distribution

Let X be a continuous random variable on $[0, N]$ such that "all points are equally likely."

This is called the uniform distribution on $[0, N]$. Its density is



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

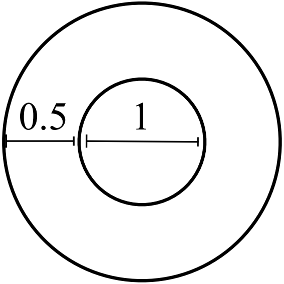
What is EX? What is VX?

Quiz Question

- Let X be a continuous random variable with density p.
- Need it be true that $p(x) < 1$?

Example: Another Uniform Distribution

Imagine that I am throwing darts on a dartboard.



Let X be the x-position of the dart I throw, and Y be the y position. Assuming that the dart is equally likely to land anywhere on the board:

1. What is the probability it will land in the inner circle?
2. What what is the joint density of X and Y?

Gaussian distribution

- The most common (and most easily analyzed) distribution for continuous quantities is the Gaussian distribution.
- Gaussian distribution is often a reasonable model for many quantities due to various central limit theorems
- Gaussian is also called the normal distribution

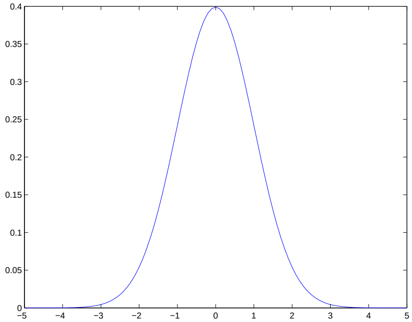
Definition

- The one-dimensional Gaussian distribution is given by

$$p(x|\mu, \sigma^2) = N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- μ is the *mean* of the Gaussian and σ^2 is the *variance*.
- If $\mu = 0$ and $\sigma^2 = 1$ then $N(x; \mu, \sigma^2)$ is called a *standard* Gaussian.

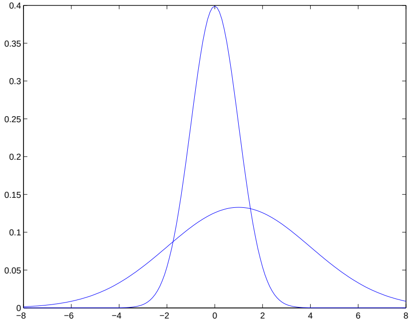
Plot



- This is a standard one dimensional Gaussian distribution.
- All Gaussians have a similar shape subject to scaling and displacement.
- If x is distributed $N(x; \mu, \sigma^2)$, then $y = (x - \mu)/\sigma$ is distributed $N(y; 0, 1)$.

Normalization

- Remember all distributions must integrate to one. The $\sqrt{2\pi\sigma^2}$ is called a normalization constant - it ensures this is the case.
- Hence tighter Gaussians have higher peaks:



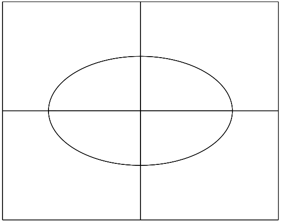
Bivariate Gaussian I

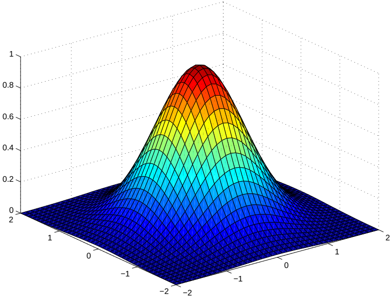
- Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$
- If X_1 and X_2 are independent

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

- Let $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$

$$p(\mathbf{x}) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\left\{(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\right\}$$




1
0.8
0.6
0.4
0.2
0
2
1
0
−1
−2
−2
−1
0
1
2

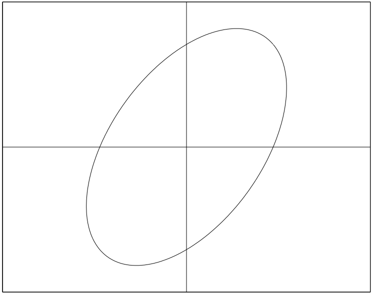
Bivariate Gaussian II

- Covariance
- Σ is the covariance matrix

$$\Sigma = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

$$\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$$

- Example: plot of weight vs height for a population



Multivariate Gaussian

- $p(\mathbf{x} \in \mathcal{R}) = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}$
- Multivariate Gaussian

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Σ is the covariance matrix

$$\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$$

$$\Sigma = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$$

- Σ is symmetric
- Shorthand $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$
- For $p(\mathbf{x})$ to be a density, Σ must be positive definite
- Σ has $d(d+1)/2$ parameters, the mean has a further d

Inverse Problem: Estimating a Distribution

- But what if we don't know the underlying distribution?
- Want to *learn* a good distribution that fits the data we do have
- How is *goodness* measured?
- Given some distribution, we can ask how likely it is to have generated the data
- In other words what is the probability (density) of this particular data set given the distribution
- A particular distribution explains the data better if the data is more probable under that distribution

Likelihood

- $p(D|M)$. The probability of the data D given a distribution (or model) M. This is called the likelihood of the model.
- This is

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

  i.e. the product of the probabilities of generating each data point individually.
- This is a result of the independence assumption.
- Try different M (different distributions). Pick the M with the highest likelihood $\rightarrow$ Maximum Likelihood Approach.

Bernoulli distribution

- Data 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 1 1 0 1, total of 20 observations
- Three hypotheses:
  - $M = 1$ - Generated from a fair coin. 1=H, 0=T
  - $M = 2$ - Generated from a die throw 1=1, 0 = 2,3,4,5,6
  - $M = 3$ - Generated from a double headed coin 1=H, 0=T
- Likelihood of data. Let c=number of ones:

$$\prod p(x_i|M) = p(1|M)^c p(0|M)^{20-c}$$

- $M = 1$: Likelihood is $0.5^{20} = 9.5 \times 10^{-7}$
- $M = 2$: Likelihood is $(1/6)^9\ (5/6)^{11} = 1.3 \times 10^{-8}$
- $M = 3$: Likelihood is $1^9\ 0^{11} = 0$

Bernoulli distribution

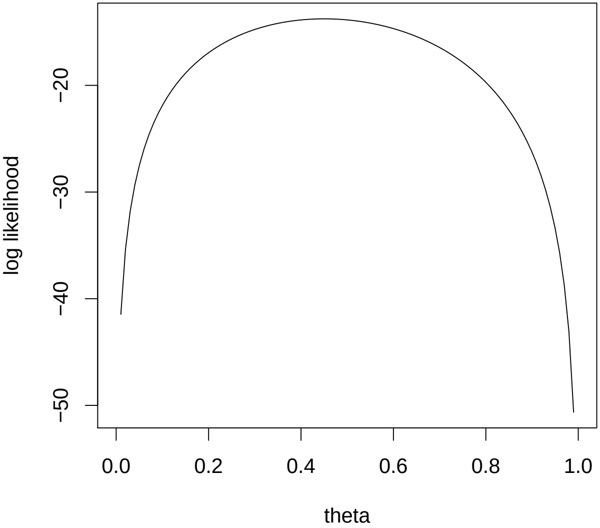
- Data 1 0 0 1 0 1 0 1 0 0 0 0 1 0 1 1 1 0 1.
- Continuous range of hypotheses: $M = \theta$ generated from a Bernoulli distribution with $p(1|M = \theta) = \theta$.
- Likelihood of data. Let $c =$number of ones in n tosses

$$\prod p(x_i|M = \theta) = \theta^c (1 - \theta)^{n-c}$$

- Maximum Likelihood hypothesis? Differentiate w.r.t. θ to find maximum
- In fact usually easier to differentiate $\log p(D|M)$: log is monotonic

$$\frac{d \log p(D|M)}{d\theta} = \frac{c}{\theta} - \frac{(n-c)}{(1-\theta)}$$

- So $c(1-\theta) - (n-c)\theta = 0$. This gives $\hat{\theta} = c/n$. Maximum likelihood result is intuitive



Notice this depends on the data set ($n = 20$, $c = 9$). With a different data set, you would get a different function of θ.

Maximum Likelihood Estimation for a Univariate Gaussian

- Suppose we have data $\{x_i, i = 1, 2, \ldots, n\}$
- Suppose we presume the data was generated from a Gaussian with mean μ and variance σ^2. Call this the model
- Then the log probability of the data given the model is

$$\log \prod_i p(x_i|\mu, \sigma^2) = -\frac{1}{2} \sum_i \frac{(x_i - \mu)^2}{\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2)$$

  Steps left as exercise: hint $\log \prod = \sum \log$
- Hence

$$\hat{\mu} = \frac{\sum_i x_i}{n}, \qquad \hat{\sigma}^2 = \frac{\sum_i (x_i - \hat{\mu})^2}{n}$$

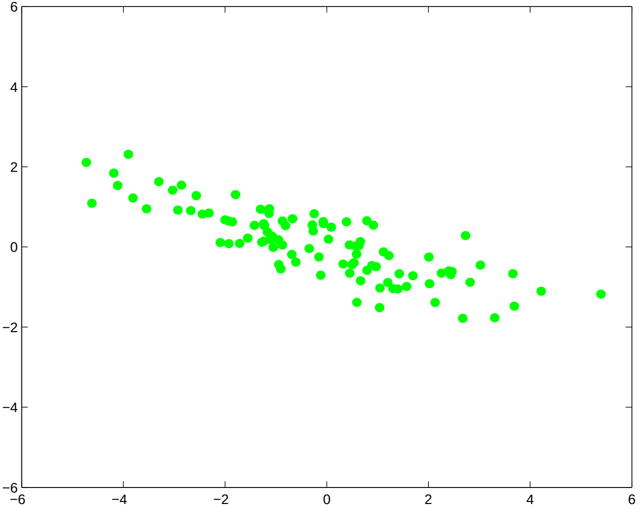
- (Maximum likelihood estimate of σ^2 is *biased*.)

Multivariate Gaussian: Maximum Likelihood

- The Maximum Likelihood estimate can be found in the same way
- $\hat{\boldsymbol{\mu}} = (1/n) \sum_{i=1}^{n} \mathbf{x}_i$
- $\hat{\Sigma} = (1/n) \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$

Example

- The data.



Example

- The data. The maximum likelihood fit.