# Probability

## Machine Learning and Pattern Recognition

Chris Williams

School of Informatics, University of Edinburgh

August 2014

(All of the slides in this course have been adapted from
previous versions by Charles Sutton, Amos Storkey, David Barber.)

# Outline

- What is probability?
- Random Variables (discrete and continuous)
- Expectation
- Joint Distributions
- Marginal Probability
- Conditional Probability
- Chain Rule
- Bayes' Rule
- Independence
- Conditional Independence
- Some Probability Distributions (for reference)
- Reading: Murphy secs 2.1-2.4

# What is probability?

- Quantification of uncertainty
- **Frequentist** interpretation: long run frequenies of events
- Example: The probability of a particular coin landing heads up is 0.43
- **Bayesian** interpretation: quantify our degrees of belief about something
- Example: the probability of it raining tomorrow is 0.3
- Not possible to repeat “tomorrow” many times
- Basic rules of probability are the same, no matter which interpretation is adopted

# Random Variables

- A random variable (RV) $X$ denotes a quantity that is subject to variations due to chance
- May denote the result of an experiment (e.g. flipping a coin) or the measurement of a real-world fluctuating quantity (e.g. temperature)
- Use capital letters to denote random variables and lower case letters to denote values that they take, e.g. $p(X = x)$
- An RV may be *discrete* or *continuous*
- A discrete variable takes on values from a finite or countably infinite set
- *Probability mass function* $p(X = x)$ for discrete random variables

- Examples:
  - 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 $X =$ (number of heads)$^2$. X can take on the values 0, 1 and 4.
- Example $p(Colour = red) = 0.3$
- $\sum_x p(x) = 1$

# Continuous RVs

- Continuous RVs take on values that vary continuously within one or more real intervals
- *Probability density function* (pdf) $p(x)$ for a continuous random variable $X$

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

  therefore

  $$p(x \leq X \leq x + \delta x) \simeq p(x)\delta x$$

- $\int p(x)dx = 1$ (but values of $p(x)$ can be greater than 1)
- Examples (coming soon): Gaussian, Gamma, Exponential, Beta

# Expectation

- Consider a function $f(x)$ mapping from $x$ onto numerical values

$$\mathbb{E}[f(x)] = \sum_x f(x)p(x)$$
$$= \int f(x)p(x)dx$$

for discrete and continuous variables resp.

- $f(x) = x$, we obtain the mean, $\mu_x$
- $f(x) = (x - \mu_x)^2$ we obtain the variance

# Joint distributions

- Properties of several random variables are important for modelling complex problems
- $p(X_1 = x_1, X_2 = x_2, \ldots, X_D = x_D)$
- "," is read as "and"
- Examples about Grade and Intelligence (from Koller and Friedman, 2009)

| | $Intelligence = low$ | $Intelligence = high$ |
|---|---|---|
| $Grade = A$ | 0.07 | 0.18 |
| $Grade = B$ | 0.28 | 0.09 |
| $Grade = C$ | 0.35 | 0.03 |

# Marginal Probability

- The *sum rule*

$$p(x) = \sum_y p(x, y)$$

- $p(Grade = A)$ ??
- Replace sum by an integral for continuous RVs

# Conditional Probability

- Let $\mathbf{X}$ and $\mathbf{Y}$ be two disjoint groups 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

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

- Product rule

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

- **Example**: In the grades example, what is $p(Intelligence = high | Grade = A)$?
- $\sum_{\mathbf{x}} p(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}) = 1 \text{ for all } \mathbf{y}$
- Can we say anything about $\sum_{\mathbf{y}} p(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y})$ ?

# Chain Rule

The chain rule is derived by repeated application of the product rule

$$
\begin{aligned}
p(X_1,\ldots,X_D) &= p(X_1,\ldots,X_{D-1})p(X_D|X_1,\ldots,X_{D-1}) \\
&= p(X_1,\ldots,X_{D-2})p(X_{D-1}|X_1,\ldots,X_{D-2}) \\
&\qquad p(X_D|X_1,\ldots,X_{D-1}) \\
&= \ldots \\
&= p(X_1)\prod_{i=2}^{D} p(X_i|X_1,\ldots,X_{i-1})
\end{aligned}
$$

- Exercise: give *six* decompositions of $p(x, y, z)$ using the chain rule

# Bayes' Rule

- From the product rule,

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

- From the sum rule the denominator is

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

# Probabilistic Inference using Bayes' Rule

- Tuberculosis (TB) and a skin test (Test)
- $p(TB = yes) = 0.001$ (for subjects who get tested)
- $p(Test = yes|TB = yes) = 0.95$
- $p(Test = no|TB = no) = 0.95$
- Person gets a positive test result. What is $p(TB = yes|Test = yes)$?

$$
\begin{aligned}
p(TB = yes|Test = yes) &= \frac{p(Test = yes|TB = yes)p(TB = yes)}{p(Test = yes)} \\
&= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.05 \times 0.999} \\
&\simeq 0.0187
\end{aligned}
$$

NB: These are fictitious numbers

# Independence

- Let $\mathbf{X}$ and $\mathbf{Y}$ be two disjoint groups of variables. Then $\mathbf{X}$ is said to be *independent* of $\mathbf{Y}$ if and only if

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

  for all possible values $\mathbf{x}$ and $\mathbf{y}$ of $\mathbf{X}$ and $\mathbf{Y}$; otherwise $\mathbf{X}$ is said to be *dependent* on $\mathbf{Y}$
- Using the definition of conditional probability, we get an equivalent expression for the independence condition

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

- $\mathbf{X}$ independent of $\mathbf{Y} \Leftrightarrow \mathbf{Y}$ independent of $\mathbf{X}$
- Independence of a set of variables. $X_1, \ldots, X_D$ are independent iff

$$p(X_1, \ldots, X_D) = \prod_{i=1}^{D} p(X_i)$$

# Conditional Independence

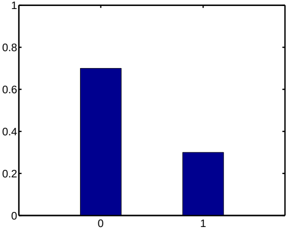
- Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be three disjoint groups of variables. $\mathbf{X}$ is said to be *conditionally independent* of $\mathbf{Y}$ given $\mathbf{Z}$ iff

  $$p(\mathbf{x}|\mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})$$

  for all possible values of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.
- Equivalently $p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})$ [show this]
- Notation, $I(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$

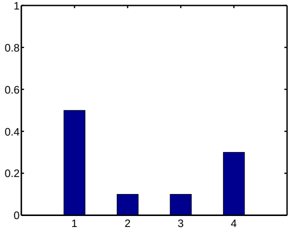
# Bernoulli Distribution

- $X$ is a random variable that either takes the value $0$ or the value $1$.
- Let $p(X = 1|p) = p$ and so $p(X = 0|p) = 1 - p$.
- Then $X$ has a Bernoulli distribution.



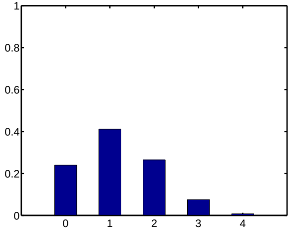
# Categorical Distribution

- $X$ is a random variable that takes one of the values $1, 2, \dots, D$.
- Let $p(X = i|\mathbf{p}) = p_i$, with $\sum_{i=1}^{D} p_i = 1$.
- Then $X$ has a catgorical (aka multinoulli) distribution (see Murphy 2012, p. 35))



# Binomial Distribution

- The binomial distribution is obtained from the total number of $1$'s in $n$ independent Bernoulli trials.
- $X$ is a random variable that takes one of the values $0, 1, 2, \ldots, n$.
- Let $p(X = r|p) = \binom{n}{r} p^r (1-p)^{(n-r)}$.
- Then $X$ is binomially distributed.



# Multinomial Distribution

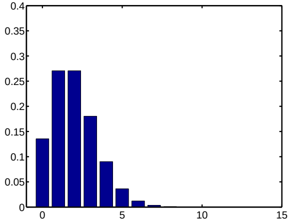
- The multinomial distribution is obtained from the total count for each outcome in $n$ independent multivariate trials with $D$ possible outcomes.
- $\boldsymbol{X}$ is a random vector of length $D$ taking values $\mathbf{x}$ with $x_i \in \mathbb{Z}^+$ (non-negative integers) and $\sum_{i=1}^{D} x_i = n$.
- Let
$$p(\boldsymbol{X} = \mathbf{x}|\mathbf{p}) = \frac{n!}{x_1! \dots x_D!} p_1^{x_1} \dots p_m^{x_D}$$
- Then $\boldsymbol{X}$ is multinomially distributed.

# Poisson Distribution

- The Poisson distribution is obtained from binomial distribution in the limit $n \to \infty$ with $p/n = \lambda$.
- $X$ is a random variable taking non-negative integer values $0, 1, 2, \ldots$.
- Let

$$p(X = x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

- Then $X$ is Poisson distributed.

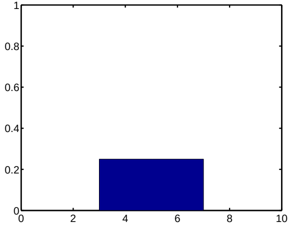


# Uniform Distribution

- $X$ is a random variable taking values $x \in [a, b]$.
- Let $p(X = x) = 1/[b - a]$
- Then $X$ is uniformly distributed.

## Note

Cannot have a uniform distribution on an unbounded region.

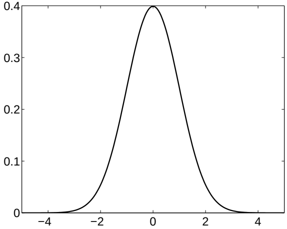


# Gaussian Distribution

- $X$ is a random variable taking values $x \in \mathbb{R}$ (real values).
- Let $p(X = x|\mu, \sigma^2) =$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

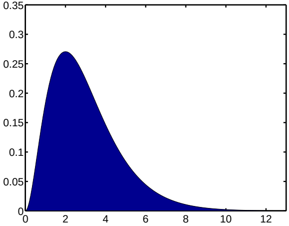
- Then $X$ is Gaussian distributed with mean $\mu$ and variance $\sigma^2$.



# Gamma Distribution

- The Gamma distribution has a rate parameter $\beta > 0$ (or a scale parameter $1/\beta$) and a shape parameter $\alpha > 0$.
- $X$ is a random variable taking values $x \in \mathbb{R}^+$ (non-negative real values).
- Let

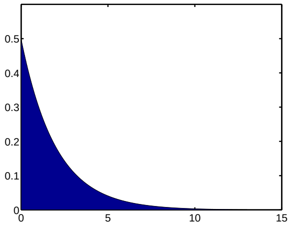
$$p(X = x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \beta^{\alpha} \exp(-\beta x)$$




- Then $X$ is Gamma distributed.
- Note the Gamma function.

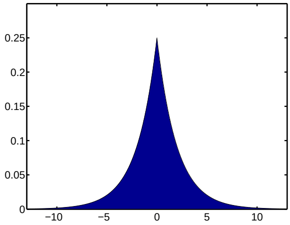
# Exponential Distribution

- The exponential distribution is a Gamma distribution with $\alpha = 1$.
- The exponential distribution is often used for arrival times.
- $X$ is a random variable taking values $x \in \mathbb{R}^+$ .
- Let $p(X = x|\lambda) = \lambda \exp(-\lambda x)$
- Then $X$ is exponentially distributed.



# Laplace Distribution

- The Laplace distribution is obtained from the difference between two independent identically exponentially distributed variables.
- $X$ is a random variable taking values $x \in \mathbb{R}$.
- Let $p(X = x|\lambda) = (\lambda/2)\exp(-\lambda|x|)$
- Then $X$ is Laplace distributed.

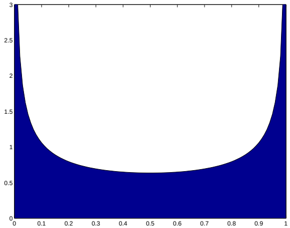


# Beta Distribution

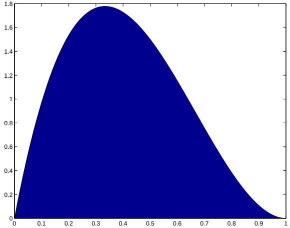
- $X$ is a random variable taking values $x \in [0, 1]$.
- Let

$$p(X = x|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1}$$

- Then $X$ is $\beta(a, b)$ distributed.




$a = b = 0.5$




$a = 2,\ b = 3$

# The Kronecker Delta

- Think of a discrete distribution with all its probability mass on one value. So $p(X = i) = 1$ iff (if and only if) $i = j$.
- We can write this using the Kronecker Delta:

$$p(X = i) = \delta_{ij}$$

- $\delta_{ij} = 1$ iff $i = j$ and is zero otherwise.

## The Dirac Delta

- Think of a real valued distribution with all its probability density on one value.
- There is an infinite density peak at one point (lets call this point $a$).
- We can write this using the Dirac delta:

$$p(X = x) = \delta(x - a)$$

which has the properties $\delta(x - a) = 0$ if $x \neq a$, $\delta(x - a) = \infty$ if $x = a$,

$$\int_{-\infty}^{\infty} dx\ \delta(x - a) = 1 \text{ and } \int_{-\infty}^{\infty} dx\ f(x)\delta(x - a) = f(a).$$

- You could think of it as a Gaussian distribution in the limit of zero variance.

# Other Distributions

- Chi-squared distribution with $k$ degrees of freedom is a Gamma distribution with $\beta = 1/2$ and $k = 2/\alpha$.
- Dirichlet distribution: will be used on this course.
- Weibull distribution (a generalisation of the exponential)
- Geometric distribution
- Negative binomial distribution.
- Wishart distribution (a distribution over matrices).
- Use Wikipedia and Mathworld. Good summaries for distributions.

# Things you must never (ever) forget

- Probabilities must be between $0$ and $1$ (though probability densities can be greater than $1$).
- Distributions must sum (or integrate) to $1$.

# Summary

- Joint distributions
- Conditional Probability
- Sum and Product Rules
- Standard Probability distributions
- Reading: Murphy secs 2.1-2.4