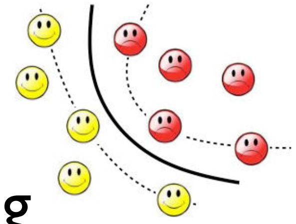
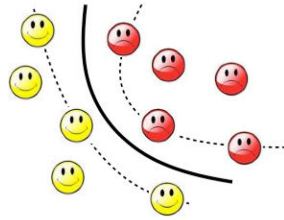


Machine Learning and Data Mining
(COMP 5318)

Basics of probability theory and Bayes' rule

Nguyen Hoang Tran



Review

Basics I

$$S = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

- This is a *3 x 3* matrix.
- In general $m \times n$.
  - m rows and n columns
  - Square matrix when $m = n$
- Each row or column could represent one object. If rows are objects then columns are features/attributes/components

Basics II

- Identity matrix I

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- If A is a square matrix, $AI = IA = A$
- I is an example of a **diagonal** matrix.
- If $A = [a_1, ...a_m]$ is matrix where a_i are the columns, then
  - A is orthogonal if $a_i.a_j = 0$ for $i \neq j$
  - A is orthonormal if above and $a_i.a_i = 1$

Basics III

- Every vector can be written as a linear combination of some finitely many "special" vectors.
- These are called **basis-vectors**.

$$S = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix} = 3\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 2\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 2\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Linear independence

- Intuitively, a set of vectors is linearly independent if any element of the set cannot be expressed as a linear combination of the others.
- The columns are not linearly independent:

$$S = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Rank of a matrix I

- Given a matrix X, the **rank** of a matrix is the maximum number of linearly independent columns.
- A rank 2 matrix: $S = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

Eigen Decomposition

- For any square matrix A we say that λ is an eigenvalue and $\mathbf{u}$ is its eigenvector if

$$A\mathbf{u} = \lambda\mathbf{u}, \quad \mathbf{u} \neq 0.$$

- Stacking up all eigenvectors/values gives

$$AU = U\Lambda = \begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

Singular Value Decomposition

- Given **any** real matrix X of size (m,n), it can be expressed as:

$$X = \underbrace{U}_{\boxed{m \times r}}\underbrace{\Sigma}_{\boxed{r \times r}}\underbrace{V^T}_{\boxed{r \times n}}$$

- r is the rank of matrix X
- U is a (m,r) column-orthonormal matrix
- V is a (n, r) column-orthonormal matrix
- Σ is diagonal r x r matrix

Example: compression of X

- Now a compact way of writing the spectral representation is:

$$A = U\Lambda U^T = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \times \mathbf{u}_i^T$$

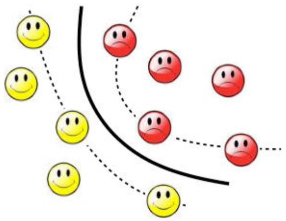
$$X = U\Sigma V^T = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \times \mathbf{v}_i^T$$

- However, can approximate it as:

$$\hat{X} = \sum_{i=1}^{k} \lambda_i \mathbf{u}_i \times \mathbf{v}_i^T$$

- This new compression ratio is:

$$\frac{mk + k + nk}{mn} = \frac{k(m+1+n)}{mn} \approx \frac{km}{mn} = \frac{k}{n} \leqslant \frac{r}{n} \approx \frac{mr + r + nr}{mn}$$



Probability Theory

Why Probabilities?

- As stated by Laplace, "Probability is common sense reduced to calculation".
- Probability theory is useful in understanding, studying, and analysis complex real world systems

Understanding uncertainty

THE UNIVERSITY OF SYDNEY

- Aleatory: chance, no ability to predict outcome
- Epistemic: encoding knowledge, ability to predict outcome
- Sensing: ability to encode noisy measurements

It is better to be imprecisely right than precisely wrong!

Predictions and Probabilities

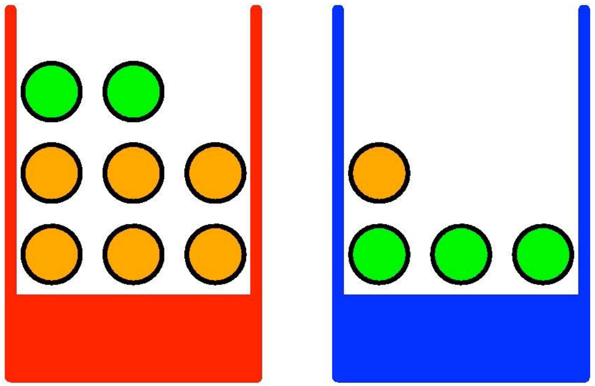
THE UNIVERSITY OF SYDNEY

- When we make predictions we should assign "probabilities" with the prediction.
- Examples:
  - 20% chance it will rain tomorrow.
  - 50% chance that the tumour is malignant.
  - 60% chance that the stock market will fall by the end of the week.
  - 30% that the next president of the United States will be a Democrat.
  - 0.1% chance that the user will click on a banner-ad.
- How do we assign probabilities to complex events… using smart data algorithms… and counting.

Probability Theory

THE UNIVERSITY OF SYDNEY

Apples and Oranges



P(apples) = 2/8 = 0.25 P(apples) = 3/4 = 0.75

Probability Basics

THE UNIVERSITY OF SYDNEY

- Probability is a deep topic.....but for most cases the rules are straightforward to apply.
- Terminology
  - Experiment
  - Sample Space
  - Events
  - Probability
  - Rules of probability
  - Conditional probability – Bayes' Rule

Experiments and Sample Space

- Consider an experiment and let S be the space of possible outcomes.
- Example:

– Experiment is tossing a coin; S={h,t}

– Experiment is rolling a pair of dice: S={(1,1),(1,2),…,(6,6)}

– Experiment is a race consisting of three cars: 1,2 and 3. The sample space is {(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,1,2),(3,2,1)}

Probability

- Let Sample Space S = {1,2,...m}
- Consider numbers $p_i \geq 0, i = 1, 2 \ldots m; \sum_i p_i = 1$
- p_i is the probability that the outcome of the experiment is i.
- Suppose we toss a fair coin. Sample space is S={h,t}. Then $p_h = 0.5$ and $p_t = 0.5$.

Assigning probabilities

- Experiment: Will it rain or not in Sydney:
  S = {rain, no-rain}
  – $P_{rain} = 138/365 = 0.38$; $P_{no\text{-}rain} = 227/365 = 0.62$
- Assigning (or rather how to obtain) probabilities is a deep philosophical problem.
  – What is the probability that the "green object standing outside my house is a burglar dressed in green?"

Events

- An *Event A* is a set of possible outcomes of the experiment. Thus A is a subset of S.
- Let A be the event of getting a seven when we roll a pair of dice.
  – A = {(1,6),(6,1),(2,5),(5,2),(4,3),(3,4) }
  – P(A) = 6/36 = 1/6
- In general $P(A) = \sum_{i \in A} p_i$

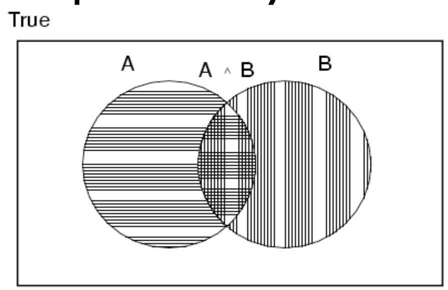
Events and Sample Space

THE UNIVERSITY OF SYDNEY

- The sample space S and events are "sets".
- $P(S) = 1$; probability of everything
- $P(\Phi) = 0$; probability of "null"
- Addition: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - Often $P(A \cap B) \equiv P(AB) \equiv P(A, B)$
- Complement: $P(A^c) = 1 - P(A)$

Axioms of probability

THE UNIVERSITY OF SYDNEY



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) \equiv P(AB) \equiv P(A, B)$$

$$P(A^c) = 1 - P(A)$$

Example

THE UNIVERSITY OF SYDNEY

- Suppose the probability of raining today is 0.4 and tomorrow is also 0.4 and on both days is 0.1. What is the probability it does not rain on either day?

Example

THE UNIVERSITY OF SYDNEY

- Suppose the probability of raining today is 0.4 and tomorrow is also 0.4 and on both days is 0.1. What is the probability it does not rain on either day?
- S={(R,N), (R,R), (N,N), (N,R)}
- Let A be the event that it will rain today and B it will rain tomorrow. Then
  A ={(R,N), (R,R)} ; B={(N,R),(R,R)}
- Rain at least today or tomorrow:
  $P(A \cup B) = 0.4 + 0.4 - 0.1 = 0.7$
- Will not rain on either day: $1 - 0.7 = 0.3$

Discrete Random Variables

- Events like "ASX is up" are binary events.
- We can extend this: by defining a **discrete random variable**.

  P(X = x) the probability that event X = x

- Two properties need to be satisfied

$$0 \leq P(X = x) \leq 1$$

$$\sum_{x \in X} P(X = x) = 1$$

P(X=x) <=1 only for discrete variables

Continuous Random Variables

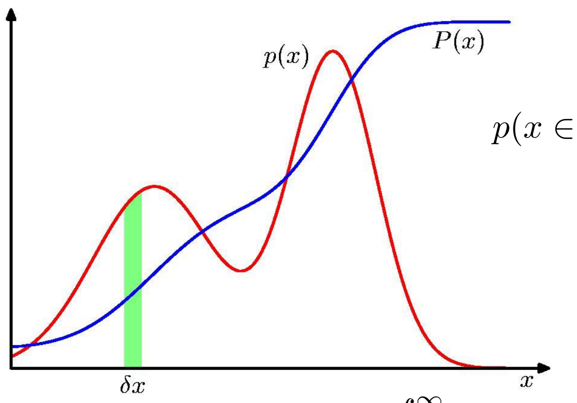
- Random variables can also be continuous:
  Height, rainfall, salary, chemical concentration…
- We can talk about the average (mean) and standard deviation or variance.
  e.g., the average height of students in COMP5318 is 175 cm with a standard deviation of 15 cm.

Probability Densities

- Random variables (both continuous and discrete) are associated with distributions.
- Common examples of discrete distributions are: Bernoulli, binomial, multinomial, Poisson.
- Common examples of continuous distributions are: Gaussian (Normal), Laplacian, Exponential, Gamma.
- Associated with distributions are parameters...
- One of the key problems in Statistics is to learn the parameters of a distribution from data.

This is **like summarising data.**

Probability Densities



probability density function (pdf)

$$p(x \in (a, b)) = \int_a^b p(x)dx$$

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

Cumulative distribution function (cdf)

$$p(x) \geq 0 \qquad \int_{-\infty}^{\infty} p(x)dx = 1$$

Expectations

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

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation (discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Approximate Expectation (discrete and continuous)

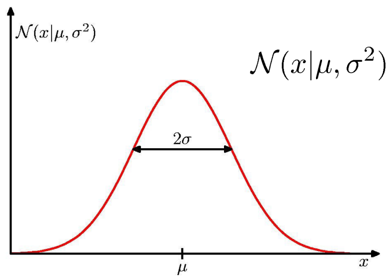
Variance and Covariance

$$\mathrm{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2 \right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

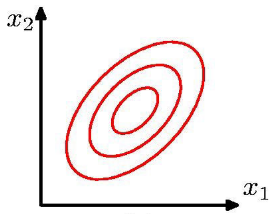
$$\begin{aligned} \mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

$$\begin{aligned} \mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}] \end{aligned}$$

The Gaussian Distribution



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



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

Most entropic distribution given a mean and variance

Gaussian Mean and Variance

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Binary Variables

Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

Bernoulli Distribution

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$$
$$\mathbb{E}[x] = \mu$$
$$\mathrm{var}[x] = \mu(1-\mu)$$

Binary Variables

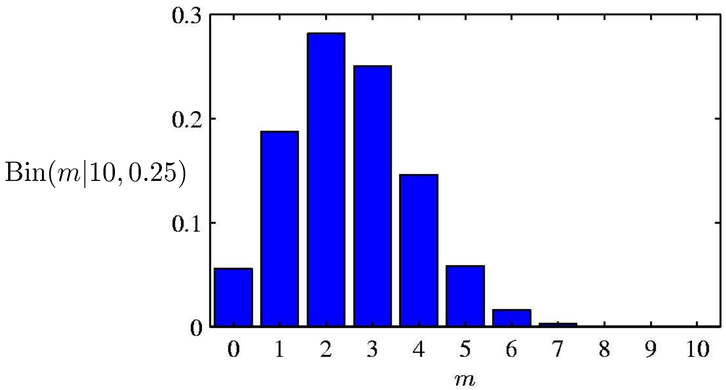
- N coin flips:

$$p(m \text{ heads}|N, \mu)$$

- Binomial Distribution

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$
$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) = N\mu$$
$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \ \mathrm{Bin}(m|N,\mu) = N\mu(1-\mu)$$

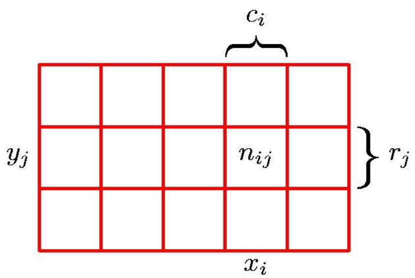
Binomial Distribution



Conditional Probabilities

- One of the most important concepts in all of Machine Learning
- P(A | B) = P(A, B)/P(B) …assuming P(B) not equal 0.
  – Conditional probability of A given B has occurred.
- Probability it will rain tomorrow given it has rained today.
  – P(A | B) = P(A, B)/P(B) = 0.1/0.4 = 1⁄4 = 0.25
  – In general P(A | B) is not equal to P(B | A)

Probability Theory



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}$$

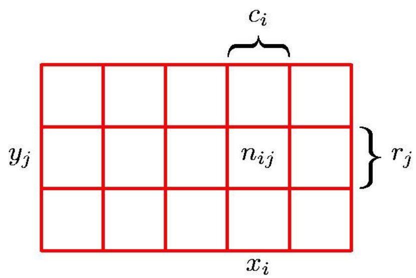
Joint Probability

$$p(X = x_i, Y = y_i) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_i | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$
$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$p(X = x_i, Y = y_i) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$
$$= p(Y = y_i | X = x_i) p(X = x_i)$$

Rules of Probability

- Sum Rule $\quad p(X) = \sum_{Y} p(X, Y)$
- Product Rule $\quad p(X, Y) = p(Y|X)p(X)$

Bayes’ Rule

- P(A | B) = P(A, B) / P(B); P(B | A) = P(B, A) | P(A)
- Now P(A, B) = P(B, A)
- Thus P(A | B) P(B) = P(B | A) P(A)
- Thus P(A | B) = [P(B | A)P(A)] / [P(B)]
  – This is called Bayes’ Rule
  – Basis of almost all prediction
  – Latest theories hypothesise that human memory and action is Bayes' rule in action.

Bayes' Rule

Likelihood

Posterior

Prior

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

Normaliser

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)}$$

Example

The ASX market goes up 60% of the days of a year. 40% of the time it stays the same or goes down. The day the ASX is up, there is a 50% chance that the Shanghai Index is up. On other days there is 30% chance that Shanghai goes up. Suppose the Shanghai market is up. What is the probability that ASX was up?

Example cont.

The ASX market goes up 60% of the days of a year. 40% of the time it stays the same or goes down. The day the ASX is up, there is a 50% chance that the Shanghai Index is up. On other days there is 30% chance that Shanghai goes up. Suppose the Shanghai market is up. What is the probability that ASX was up?

- We want to calculate P(A1 | S1) ?
- P(A1) = 0.6; P(A2) = 0.4;
  P(S1 | A1) = 0.5; P(S1 | A2) = 0.3
  P(S2 | A1) = 1 – P(S1 | A1) = 0.5;
  P(S2 | A2) = 1 – P(S1 | A2) = 0.7;
- P(A1 | S1) = P(S1 | A1)P(A1) / (P(S1))
- How do we calculate P(S1) ?

Example cont.

- P(S1) = P(S1,A1) + P(S1,A2) [Key Step]
  = P(S1 | A1)P(A1) + P(S1 | A2)P(A2)
  = 0.5 x 0.6 + 0.3 x 0.4
  = 0.42
- Finally,
  P(A1 | S1) = P(S1 | A1)P(A1) / P(S1)
  = (0.5 x 0.6)/0.42
  = 0.71

Independence

- Two events A and B are independent if

$$P(A,B) = P(A)P(B)$$

- Example: Toss a coin twice. Then what is the probability of two heads?
- The outcome of the two tosses are not dependent on each other

$$P(H,H) = P(H)P(H) = 0.5 \times 0.5 = 0.25$$

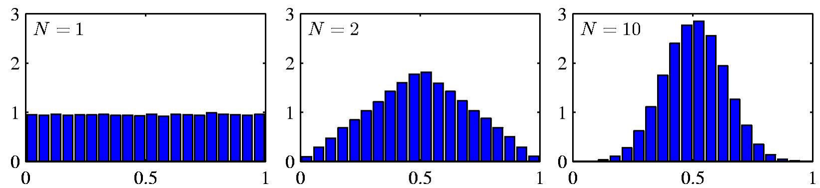
- If A and B are independent then

$$P(A \mid B) = P(A,B) / P(B) = P(A)P(B) / P(B) = P(A) \; !$$

Central Limit Theorem

The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows.

Example: N uniform [0,1] random variables.



Expectations

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

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x)$$ Conditional Expectation (discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$ Approximate Expectation (discrete and continuous)

The law of large numbers

LLN describes the result of performing the same experiment a large number of times.

The average of the results obtained from a large number of independent trials should converge to the expected value.

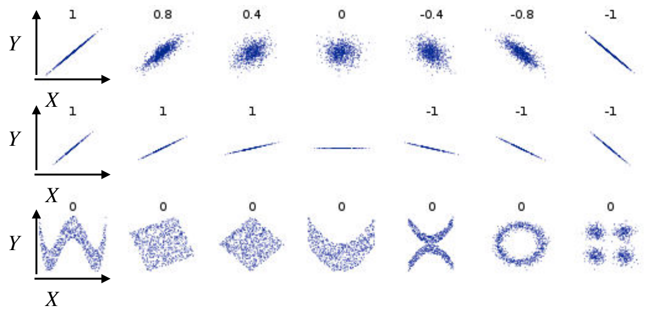
$$\frac{\sum_{I=1}^{n} 1_{\{x_i = \text{“}head\text{”}\}}}{n} \rightarrow \int P(x = \text{“}head\text{”}) 1_{\{x = \text{“}head\text{”}\}} dx$$

$$= P(x = \text{“}head\text{”})$$

Correlation vs dependence

Correlation coefficient:

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$



Even though the correlation coef is zero they are still dependent!