UNIVERSITY
of
GLASGOW

Machine Learning

Lecture. 4.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

Math & Probability Basics

- Some of the basic maths and probability required for Week 3 & 4 material

Math & Probability Basics

UNIVERSITY of GLASGOW

- Some of the basic maths and probability required for Week 3 & 4 material
- Linear Algebra basics

Math & Probability Basics

UNIVERSITY
of
GLASGOW

- Some of the basic maths and probability required for Week 3 & 4 material
- Linear Algebra basics
- Probability & Probability Distributions

Math & Probability Basics

UNIVERSITY *of* GLASGOW

A D-dimensional column vector defined as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_D \end{bmatrix}$$

Math & Probability Basics

UNIVERSITY
of
GLASGOW

A D-dimensional row vector defined as transpose of D-dimensional column vector

$$\mathbf{x}^{\mathsf{T}} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_D \end{bmatrix}$$

Math & Probability Basics

UNIVERSITY of GLASGOW

Inner product of two vectors $\mathbf{a}^\mathsf{T}\mathbf{b}$ defined as

$$\begin{aligned}
\mathbf{a}^\mathsf{T}\mathbf{b} &= \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_D \end{bmatrix} \times \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_D \end{bmatrix} \\
&= a_1 b_1 + a_2 b_2 + a_3 b_3 + \cdots + a_D b_D = \sum_{i=1}^{D} a_i b_i
\end{aligned}$$

Math & Probability Basics

UNIVERSITY *of* GLASGOW

Euclidean norm or length of vector

$$||\mathbf{x}|| = \sqrt{\mathbf{x}^\mathsf{T}\mathbf{x}}$$

Vector has unit norm if $||\mathbf{x}|| = 1$
The angle θ between two vectors $\mathbf{a}$ and $\mathbf{b}$ defined by

$$\cos(\theta) = \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{||\mathbf{a}||||\mathbf{b}||}$$

If $\cos(\theta) = 0$, i.e. $\mathbf{a}^\mathsf{T}\mathbf{b} = 0$ then vectors are orthogonal

Math & Probability Basics

UNIVERSITY *of* GLASGOW

A set of N D-dimensional vectors $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ are linearly independent if no vector in the set can be written as linear combination of any of the others.

A set of N linearly independent vectors span an N-dimensional vector space

Any vector in this space can be represented by a linear combination of these basis vectors. Basis in 3-D space

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \; \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \; \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Outer Product

The outer-product of an N-dimensional vector $\mathbf{a}$ and an M-dimensional vector $\mathbf{b}$ defined as

$$\mathbf{a}\mathbf{b}^{\mathsf{T}} = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_M \\ a_2 b_1 & a_2 b_2 & \cdots & a_2 b_M \\ \vdots & \cdots & \cdots & \vdots \\ a_N b_1 & a_N b_2 & \cdots & a_n b_M \end{bmatrix}$$

Matrix Derivatives

UNIVERSITY *of* GLASGOW

A scalar function of a D-dimensional vector $\mathbf{x}$ defined as $f(\mathbf{x})$ then the derivative of $f(\mathbf{x})$ with respect to $\mathbf{x}$ is defined as

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_D} \end{bmatrix}$$

Matrix Derivatives

For example if $f(\mathbf{x}) = \mathbf{a}^{\mathsf{T}}\mathbf{x}$ then

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{a}^{\mathsf{T}}\mathbf{x} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_D \end{bmatrix} = \mathbf{a}$$

Matrix Derivatives

UNIVERSITY *of* GLASGOW

For a N-dimensional vector valued function $\mathbf{f}(\mathbf{x})$, where $\mathbf{x}$ is D-dimensional the Jacobian matrix is defined as

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \frac{\partial \mathbf{f}_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}_1(\mathbf{x})}{\partial x_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{f}_N(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}_N(\mathbf{x})}{\partial x_D} \end{bmatrix}$$

Matrix Derivatives

UNIVERSITY *of* GLASGOW

Lets say we have a function $f(\mathbf{x}) = \left(\mathbf{a}^\mathsf{T}\mathbf{x}\right)^2$ then

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} 2\mathbf{a}^\mathsf{T}\mathbf{x}a_1 \\ 2\mathbf{a}^\mathsf{T}\mathbf{x}a_2 \\ \vdots \\ 2\mathbf{a}^\mathsf{T}\mathbf{x}a_D \end{bmatrix}$$

Matrix Derivatives

UNIVERSITY *of* GLASGOW

Now we can take the second partial derivatives

$$\frac{\partial}{\partial \mathbf{x}}\left(\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x})\right) = \frac{\partial}{\partial \mathbf{x}} \begin{bmatrix} 2\mathbf{a}^{\mathsf{T}}\mathbf{x}a_1 \\ 2\mathbf{a}^{\mathsf{T}}\mathbf{x}a_2 \\ \vdots \\ 2\mathbf{a}^{\mathsf{T}}\mathbf{x}a_D \end{bmatrix}$$

$$= 2 \begin{bmatrix} a_1^2 & a_2 a_1 & \cdots & a_D a_1 \\ a_1 a_2 & a_2^2 & \cdots & a_D a_2 \\ \vdots & \vdots & \vdots & \vdots \\ a_1 a_D & a_2 a_D & \cdots & a_D^2 \end{bmatrix} = 2\mathbf{a}\mathbf{a}^{\mathsf{T}}$$

Matrix Identities

UNIVERSITY *of* GLASGOW

- The Determinant of a square $N \times N$ matrix $\mathbf{M}$ denoted as $\det(\mathbf{M})$ or $|\mathbf{M}|$provides useful information

Matrix Identities

UNIVERSITY *of* GLASGOW

- The Determinant of a square $N \times N$ matrix $\mathbf{M}$ denoted as $\det(\mathbf{M})$ or $|\mathbf{M}|$provides useful information
- If columns of $\mathbf{M}$ are not linearly independent then $\det(\mathbf{M}) = 0$, indicating that rank of matrix $\mathbf{M}$ is smaller than N and $\mathbf{M}$ is not uniquely invertible

Matrix Identities

UNIVERSITY *of* GLASGOW

- The Determinant of a square $N \times N$ matrix $\mathbf{M}$ denoted as $\det(\mathbf{M})$ or $|\mathbf{M}|$provides useful information
- If columns of $\mathbf{M}$ are not linearly independent then $\det(\mathbf{M}) = 0$, indicating that rank of matrix $\mathbf{M}$ is smaller than N and $\mathbf{M}$ is not uniquely invertible
- $\det(\mathbf{M})$ is a measure of the volume deformation when $\mathbf{M}$ is used as a linear transformation, large values indicating large amounts of stretching

Matrix Identities

UNIVERSITY *of* GLASGOW

- The Determinant of a square $N \times N$ matrix $\mathbf{M}$ denoted as $\det(\mathbf{M})$ or $|\mathbf{M}|$provides useful information
- If columns of $\mathbf{M}$ are not linearly independent then $\det(\mathbf{M}) = 0$, indicating that rank of matrix $\mathbf{M}$ is smaller than N and $\mathbf{M}$ is not uniquely invertible
- $\det(\mathbf{M})$ is a measure of the volume deformation when $\mathbf{M}$ is used as a linear transformation, large values indicating large amounts of stretching
- $\det(\mathbf{M}) = \prod_{n=1}^{N} \lambda_n$ where each λ_n are the eigenvalues of $\mathbf{M}$. (more on eigenvalues later)

Matrix Identities

UNIVERSITY *of* GLASGOW

- The Determinant of a square $N \times N$ matrix $\mathbf{M}$ denoted as $\det(\mathbf{M})$ or $|\mathbf{M}|$provides useful information
- If columns of $\mathbf{M}$ are not linearly independent then $\det(\mathbf{M}) = 0$, indicating that rank of matrix $\mathbf{M}$ is smaller than N and $\mathbf{M}$ is not uniquely invertible
- $\det(\mathbf{M})$ is a measure of the volume deformation when $\mathbf{M}$ is used as a linear transformation, large values indicating large amounts of stretching
- $\det(\mathbf{M}) = \prod_{n=1}^{N} \lambda_n$ where each λ_n are the eigenvalues of $\mathbf{M}$. (more on eigenvalues later)
- The trace of a matrix is the sum of its diagonal elements $\mathrm{trace}(\mathbf{M}) = \sum_{n=1}^{N} M_{nn}$

Matrix Identities

UNIVERSITY *of* GLASGOW

- If the determinant of the square matrix $\mathbf{M}$ is non-zero then the inverse is denoted as $\mathbf{M}^{-1}$ and $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix

Matrix Identities

UNIVERSITY *of* GLASGOW

- If the determinant of the square matrix $\mathbf{M}$ is non-zero then the inverse is denoted as $\mathbf{M}^{-1}$ and $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix
- if $\mathbf{M}$ is non-square then the pseudo-inverse is given as $\mathbf{M}^{\dagger} = \left(\mathbf{M}^{\mathsf{T}}\mathbf{M}\right)^{-1}\mathbf{M}^{\mathsf{T}}$ and so $\mathbf{M}^{\dagger}\mathbf{M} = \mathbf{I}$.

Matrix Identities

UNIVERSITY *of* GLASGOW

- An important class of linear equations take the form $\mathbf{M}\mathbf{x} = \lambda\mathbf{x}$ in other words applying a transformation $\mathbf{M}$ to the vector $\mathbf{x}$ simply amounts to a scaling by λ

Matrix Identities

UNIVERSITY *of* GLASGOW

- An important class of linear equations take the form $\mathbf{Mx} = \lambda\mathbf{x}$ in other words applying a transformation $\mathbf{M}$ to the vector $\mathbf{x}$ simply amounts to a scaling by λ
- Solving for $\mathbf{x}$ and λ requires $(\mathbf{M} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$

Matrix Identities

UNIVERSITY *of* GLASGOW

- An important class of linear equations take the form $\mathbf{M}\mathbf{x} = \lambda\mathbf{x}$ in other words applying a transformation $\mathbf{M}$ to the vector $\mathbf{x}$ simply amounts to a scaling by λ
- Solving for $\mathbf{x}$ and λ requires $(\mathbf{M} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$
- For $\mathbf{M}$ real and symmetric there are N solution (eigen) vectors $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N\}$ and corresponding coefficients (eigenvalues) $\{\lambda_1, \lambda_2, \cdots, \lambda_N\}$ such that $\mathbf{e}_i^{\mathsf{T}}\mathbf{e}_j = \delta_{ij}$ if $\lambda_i \neq \lambda_j$

Matrix Identities

UNIVERSITY of GLASGOW

- An important class of linear equations take the form $\mathbf{M}\mathbf{x} = \lambda\mathbf{x}$ in other words applying a transformation $\mathbf{M}$ to the vector $\mathbf{x}$ simply amounts to a scaling by λ
- Solving for $\mathbf{x}$ and λ requires $(\mathbf{M} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$
- For $\mathbf{M}$ real and symmetric there are N solution (eigen) vectors $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N\}$ and corresponding coefficients (eigenvalues) $\{\lambda_1, \lambda_2, \cdots, \lambda_N\}$ such that $\mathbf{e}_i^{\mathsf{T}}\mathbf{e}_j = \delta_{ij}$ if $\lambda_i \neq \lambda_j$
- Eigenvectors form a basis of the N-dimensional space so transformation by $\mathbf{M}$ performs scaling of λ_i along each axis

Probability

UNIVERSITY of GLASGOW

- Let X be a discrete random variable that can take on any of D values from the set $\mathcal{X} = \{v_1, v_2, \cdots, v_D\}$

Probability

UNIVERSITY of GLASGOW

- Let X be a discrete random variable that can take on any of D values from the set $\mathcal{X} = \{v_1, v_2, \cdots, v_D\}$
- Probability that X takes value v_i denoted as $p_i = Pr(X = v_i) = P(x_i)$ for $i = 1, \cdots, D$, known as Probability Mass Function

Probability

UNIVERSITY of GLASGOW

- Let X be a discrete random variable that can take on any of D values from the set $\mathcal{X} = \{v_1, v_2, \cdots, v_D\}$
- Probability that X takes value v_i denoted as $p_i = Pr(X = v_i) = P(x_i)$ for $i = 1, \cdots, D$, known as Probability Mass Function
- Probabilities p_i must satisfy conditions $p_i \geq 0$ and $\sum_{i=1}^{D} p_i = 1$

Probability

UNIVERSITY of GLASGOW

- Rules of Probability for Discrete Variables

Probability

UNIVERSITY
of
GLASGOW

- Rules of Probability for Discrete Variables
- Two Variables $X \in \mathcal{X}$ & $Y \in \mathcal{Y}$

Probability

UNIVERSITY *of* GLASGOW

- Rules of Probability for Discrete Variables
- Two Variables $X \in \mathcal{X}$ & $Y \in \mathcal{Y}$
- Probability $P(x, y) \geq 0$ and $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$

Probability

UNIVERSITY of GLASGOW

- Rules of Probability for Discrete Variables
- Two Variables $X \in \mathcal{X}$ & $Y \in \mathcal{Y}$
- Probability $P(x, y) \geq 0$ and $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$
- Then $P(x) = \sum_{y \in \mathcal{Y}} P(x, y)$ and $P(y) = \sum_{x \in \mathcal{X}} P(x, y)$

Probability

UNIVERSITY of GLASGOW

- Rules of Probability for Discrete Variables
- Two Variables $X \in \mathcal{X}$ & $Y \in \mathcal{Y}$
- Probability $P(x, y) \geq 0$ and $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$
- Then $P(x) = \sum_{y \in \mathcal{Y}} P(x, y)$ and $P(y) = \sum_{x \in \mathcal{X}} P(x, y)$
- $P(x, y) = P(x|y)P(y) = P(y|x)P(x)$

Probability

UNIVERSITY *of* GLASGOW

- Rules of Probability for Discrete Variables
- Two Variables $X \in \mathcal{X}$ & $Y \in \mathcal{Y}$
- Probability $P(x,y) \geq 0$ and $\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(x,y) = 1$
- Then $P(x) = \sum_{y\in\mathcal{Y}} P(x,y)$ and $P(y) = \sum_{x\in\mathcal{X}} P(x,y)$
- $P(x,y) = P(x|y)P(y) = P(y|x)P(x)$
- Bayes Rule

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x\in\mathcal{X}} P(x,y)} = \frac{P(y|x)P(x)}{\sum_{x\in\mathcal{X}} P(y|x)P(x)}$$

Probability

- The expected value (mean, average) of the random variable X is $E\{X\} = \mu = \sum_{i=1}^{D} v_i p_i = \sum_{x \in \mathcal{X}} x P(x)$

Probability

UNIVERSITY *of* GLASGOW

- The expected value (mean, average) of the random variable X is $E\{X\} = \mu = \sum_{i=1}^{D} v_i p_i = \sum_{x \in \mathcal{X}} x P(x)$
- More generally $E\{f(X)\} = \sum_{i=1}^{D} f(v_i) p_i = \sum_{x \in \mathcal{X}} f(x) P(x)$
- Now variance defined as

$$\begin{aligned} \sigma^2 = E\{(X-\mu)^2\} &= \sum_{x \in \mathcal{X}} (x-\mu)^2 P(x) \\ &= E\{X^2\} - (E\{X\})^2 \end{aligned}$$

Probability

UNIVERSITY
of
GLASGOW

- Continuous random variables - cannot think of X taking on a particular value

Probability

- Continuous random variables - cannot think of X taking on a particular value
- Think of probability that value of $X = x$ falls in some range $[a, b]$

Probability

UNIVERSITY *of* GLASGOW

- Continuous random variables - cannot think of X taking on a particular value
- Think of probability that value of $X = x$ falls in some range $[a, b]$
- No longer have probability mass function $P(X = x)$ - now probability density function $p(X = x)$ use $p(x)$ as shorthand

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

Probability

UNIVERSITY *of* GLASGOW

- Continuous random variables - cannot think of X taking on a particular value
- Think of probability that value of $X = x$ falls in some range $[a, b]$
- No longer have probability mass function $P(X = x)$ - now probability density function $p(X = x)$ use $p(x)$ as shorthand

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

- Density function must satisfy $p(x) \geq 0$ and $\int_{-\infty}^{+\infty} p(x)dx = 1$

Probability

UNIVERSITY of GLASGOW

- Expectations follow as before

$$E\{X\} = \mu = \int_{-\infty}^{+\infty} xp(x)dx$$

and

$$\begin{aligned} \sigma^2 = E\{(X-\mu)^2\} &= \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx \\ &= E\{X^2\} - \mu^2 \end{aligned}$$

Probability

UNIVERSITY *of* GLASGOW

- Important probability density function is Gaussian or Normal

Probability

UNIVERSITY *of* GLASGOW

- Important probability density function is Gaussian or Normal
- Defined for single variable as

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

Probability

UNIVERSITY of GLASGOW

- Important probability density function is Gaussian or Normal
- Defined for single variable as

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

- Denoted as $p(x) = \mathcal{N}_x(\mu, \sigma)$ in class notes

Probability

UNIVERSITY
of
GLASGOW

- What about multiple variables e.g. $X_1, X_2, \cdots, X_D$

Probability

UNIVERSITY of GLASGOW

- What about multiple variables e.g. $X_1, X_2, \cdots, X_D$
- Follows from results for discrete variables (exchange integrals for summations)

Probability

UNIVERSITY of GLASGOW

- What about multiple variables e.g. $X_1, X_2, \cdots, X_D$
- Follows from results for discrete variables (exchange integrals for summations)
- Define $p(x_1, x_2, \cdots, x_D) = p(\mathbf{x}) \geq 0$ and

$$\int_{x_1=-\infty}^{x_1=+\infty} \cdots \int_{x_D=-\infty}^{x_D=+\infty} p(x_1, x_2, \cdots, x_D) dx_1 dx_2 \cdots dx_D$$

$$\equiv \int p(\mathbf{x}) d\mathbf{x} = 1$$

Probability

UNIVERSITY of GLASGOW

- Consider case of two variables x and y joint probability is $p(x, y)$ and can be decomposed as

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

Probability

UNIVERSITY of GLASGOW

- Consider case of two variables x and y joint probability is $p(x, y)$ and can be decomposed as

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

- If x and y are independent then probability of x will not be conditional upon y, $p(x|y) = p(x)$ and the probability of y will not be conditional upon x, i.e. $p(y|x) = p(y)$, so $p(x, y) = p(x)p(y)$

Probability

UNIVERSITY *of* GLASGOW

- Consider case of two variables x and y joint probability is $p(x, y)$ and can be decomposed as

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

- If x and y are independent then probability of x will not be conditional upon y, $p(x|y) = p(x)$ and the probability of y will not be conditional upon x, i.e. $p(y|x) = p(y)$, so $p(x, y) = p(x)p(y)$
- General case if all variables are independent then

$$p(\mathbf{x}) = p(x_1, x_2, \cdots, x_D) = \prod_{d=1}^{D} p(x_d)$$

Probability

UNIVERSITY *of* GLASGOW

- Back to two variables x and y joint probability is $p(x, y)$

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

Probability

- Back to two variables x and y joint probability is $p(x, y)$

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

- So Bayes Theorem gives

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \quad p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Probability

UNIVERSITY of GLASGOW

- Back to two variables x and y joint probability is $p(x, y)$

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

- So Bayes Theorem gives

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \quad p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- and

$$\begin{aligned} p(x) &= \int p(x, y)dy = \int p(x|y)p(y)dy \\ p(y) &= \int p(x, y)dx = \int p(y|x)p(x)dx \end{aligned}$$

Probability

UNIVERSITY
of
GLASGOW

- First and Second Moments defined for random vector

Probability

UNIVERSITY *of* GLASGOW

- First and Second Moments defined for random vector
- First Moment (Mean Vector) defined as

$$E\{\mathbf{x}\} = \boldsymbol{\mu} = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

Probability

- First and Second Moments defined for random vector
- First Moment (Mean Vector) defined as

$$E\{\mathbf{x}\} = \boldsymbol{\mu} = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

- Second Moment (Covariance Matrix) multivariate generalisation of variance

$$\begin{aligned} \boldsymbol{\Sigma} &= \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x} \\ &= E\left\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\right\} \end{aligned}$$

Probability

Covariance matrix has form

$$\mathbf{\Sigma} = \begin{bmatrix} E\{(x_1-\mu_1)(x_1-\mu_1)\} & \cdots & E\{(x_1-\mu_1)(x_D-\mu_D)\} \\ E\{(x_2-\mu_2)(x_1-\mu_1)\} & \cdots & E\{(x_2-\mu_2)(x_D-\mu_D)\} \\ \vdots & \ddots & \vdots \\ E\{(x_D-\mu_D)(x_1-\mu_1)\} & \cdots & E\{(x_D-\mu_D)(x_D-\mu_D)\} \end{bmatrix}$$

Probability

Covariance matrix has form

$$\mathbf{\Sigma} = \begin{bmatrix} E\{(x_1-\mu_1)(x_1-\mu_1)\} & \cdots & E\{(x_1-\mu_1)(x_D-\mu_D)\} \\ E\{(x_2-\mu_2)(x_1-\mu_1)\} & \cdots & E\{(x_2-\mu_2)(x_D-\mu_D)\} \\ \vdots & \ddots & \vdots \\ E\{(x_D-\mu_D)(x_1-\mu_1)\} & \cdots & E\{(x_D-\mu_D)(x_D-\mu_D)\} \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D1} & \sigma_{D2} & \cdots & \sigma_D^2 \end{bmatrix}$$

Probability

UNIVERSITY of GLASGOW

- Multivariate Gaussian density function

Probability

UNIVERSITY *of* GLASGOW

- Multivariate Gaussian density function
- Assume that D random variables are independent and each has a Gaussian distribution $p(x_d) = \mathcal{N}_{x_d}(\mu_d, \sigma_d)$

Probability

UNIVERSITY *of* GLASGOW

- Multivariate Gaussian density function
- Assume that D random variables are independent and each has a Gaussian distribution $p(x_d) = \mathcal{N}_{x_d}(\mu_d, \sigma_d)$
- $p(x_1, \cdots, x_D) = p(\mathbf{x}) = \prod_{d=1}^{D} p(x_d) = \prod_{d=1}^{D} \mathcal{N}_{x_d}(\mu_d, \sigma_d)$

Probability

UNIVERSITY *of* GLASGOW

- Multivariate Gaussian density function
- Assume that D random variables are independent and each has a Gaussian distribution $p(x_d) = \mathcal{N}_{x_d}(\mu_d, \sigma_d)$
- $p(x_1, \cdots, x_D) = p(\mathbf{x}) = \prod_{d=1}^D p(x_d) = \prod_{d=1}^D \mathcal{N}_{x_d}(\mu_d, \sigma_d)$
- and

$$
\begin{aligned}
p(\mathbf{x}) &= \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left\{-\frac{1}{2\sigma_d^2}(x_d - \mu_d)^2\right\} \\
&= \frac{1}{2\pi^{\frac{D}{2}} \prod_{d=1}^D \sigma_d} \exp\left\{-\frac{1}{2}\sum_{d=1}^D \left(\frac{x_d - \mu_d}{\sigma_d}\right)^2\right\}
\end{aligned}
$$

Probability

UNIVERSITY *of* GLASGOW

- Define covariance matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D^2 \end{bmatrix}$$

Probability

UNIVERSITY *of* GLASGOW

- Define covariance matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D^2 \end{bmatrix}$$

- So inverse of covariance matrix $\mathbf{\Sigma}^{-1}$ is simply

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_D^2} \end{bmatrix}$$

Probability

UNIVERSITY of GLASGOW

- Using vector notation

$$\sum_{d=1}^{D} \left(\frac{x_d - \mu_d}{\sigma_d} \right)^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Probability

UNIVERSITY *of* GLASGOW

- Using vector notation

$$\sum_{d=1}^{D} \left(\frac{x_d - \mu_d}{\sigma_d} \right)^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- Now for a diagonal matrix $\boldsymbol{\Sigma}$ then $\prod_{d=1}^{D} \sigma_d = \det(\boldsymbol{\Sigma})$

Probability

UNIVERSITY of GLASGOW

- Using vector notation

$$\sum_{d=1}^{D} \left(\frac{x_d - \mu_d}{\sigma_d} \right)^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- Now for a diagonal matrix $\mathbf{\Sigma}$ then $\prod_{d=1}^{D} \sigma_d = \det(\mathbf{\Sigma})$
- The general form for a multivariate Gaussian follows as

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

- This is the general form which holds even if $\mathbf{\Sigma}$ is not diagonal.

Probability

UNIVERSITY of GLASGOW

- Using vector notation

$$\sum_{d=1}^{D} \left(\frac{x_d - \mu_d}{\sigma_d} \right)^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- Now for a diagonal matrix $\boldsymbol{\Sigma}$ then $\prod_{d=1}^{D} \sigma_d = \det(\boldsymbol{\Sigma})$
- The general form for a multivariate Gaussian follows as

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

- This is the general form which holds even if $\boldsymbol{\Sigma}$ is not diagonal.