week 3-1 20/June/2018

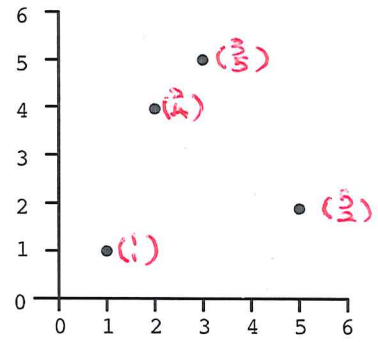
Linear algebra I

Basic vector-matrix notation, and dot products

Topics we'll cover

1. Representing data using vectors and matrices
2. Vector and matrix notation
3. Taking the transpose
4. Dot products, angles, and orthogonality

Data as vectors and matrices



- Represent the entire dataset in a single matrix.

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \\ 3 & 5 \\ 5 & 2 \end{bmatrix}$$ 4×2 (row×col) matrix

$$\begin{bmatrix} 1 & 2 & 3 & 5 \\ 1 & 4 & 5 & 2 \end{bmatrix}$$ 2×4 matrix

Matrix-vector notation

Vector $x \in \mathbb{R}^d$:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{pmatrix}$$

d×1 vector.
(d rows, 1 col)

Matrix $M \in \mathbb{R}^{r \times d}$:

$$M = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1d} \\ M_{21} & M_{22} & \cdots & M_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ M_{r1} & M_{r2} & \cdots & M_{rd} \end{pmatrix}$$

r (rows)

d (columns)

M_{ij} = entry at row i, column j

r×d matrix

Transpose of vectors and matrices

- Switch the rows and columns.

$$x = \begin{pmatrix} 1 \\ 6 \\ 3 \\ 0 \end{pmatrix} \text{ has transpose } x^T = \begin{pmatrix} 1 & 6 & 3 & 0 \end{pmatrix}$$

4×1 $\quad$ 1×4

$$M = \begin{pmatrix} 1 & 2 & 0 & 4 \\ 3 & 9 & 1 & 6 \\ 8 & 7 & 0 & 2 \end{pmatrix} \text{ has transpose } M^T = \begin{pmatrix} 1 & 3 & 8 \\ 2 & 9 & 7 \\ 0 & 1 & 0 \\ 4 & 6 & 2 \end{pmatrix}$$

3×4 $\quad$ 4×3 matrix

- $(A^T)_{ij} = A_{ji}$ → rows and cols are switched.
- $(A^T)^T = A$

double transport = original matrix

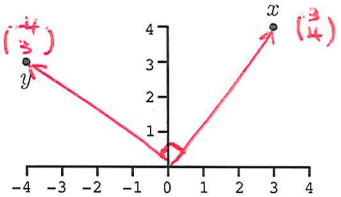
Dot product of two vectors

- 2 vectors in d-dimensions. (each have d-coordinates)

Dot product of vectors $x, y \in \mathbb{R}^d$:

$$x \cdot y = x_1 y_1 + x_2 y_2 + \cdots + x_d y_d.$$

What is the dot product between these two vectors?



$\begin{pmatrix} 3 \\ 4 \end{pmatrix} \cdot \begin{pmatrix} -4 \\ 3 \end{pmatrix} = 3 \times -4 + 4 \times 3$
$= 0.$

Adding and subtracting vectors and matrices

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 5 \\ 7 \\ 9 \end{pmatrix}$$

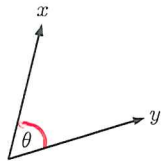
$$\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 3 \\ 3 & 6 \end{pmatrix}$$

Dot products and angles

- dot product between the vectors ~ tells the angles between them.

Dot product of vectors $x, y \in \mathbb{R}^d$: $x \cdot y = x_1 y_1 + x_2 y_2 + \cdots + x_d y_d$.

Tells us the angle between x and y:



$$\cos\theta = \frac{x \cdot y}{\|x\| \, \|y\|}$$

→ dot product of x and y
→ length of x and length y.

$x \cdot y = 0 \Rightarrow \cos\theta = 0 \Rightarrow \theta = 90°$

x is **orthogonal** (at right angles) to y if and only if $x \cdot y = 0$ → To see if two things are right angles? Just check $x \cdot y$ if it equals to 0.

When x, y are **unit vectors** (length 1): $\cos\theta = x \cdot y$

$\|x\|\|y\|$ - denominator

What is $x \cdot x$?

$x, x \in \mathbb{R}^d$: $x \cdot x = x_1^2 + x_2^2 + x_3^2 \cdots x_d^2$
$= \|x\|^2$

Another way using vectors ~ indicating directions

- if we're merely indicating directions the length doesn't matter.
- Commonly, normalize vector to length 1.

Angle between x and itself $= 0 \Rightarrow \cos 0 = 1$

$$1 = \cos 0 = \frac{x \cdot x}{\|x\| \|x\|} = \frac{\|x\|^2}{\|x\| \|x\|} = 1$$

Week 3-2. 20/June/2018

Linear algebra II
Linear functions and matrix products

Topics we'll cover

- 3.1. Dot Product between 2 vectors
- Now. Products between vectors / matrix and matrix-matrix

1. Linear functions
2. Matrix-vector products
3. Matrix-matrix products

Linear and quadratic functions

In one dimension:

- Linear: $f(x) = 3x + 2$ $= ax + b$
- Quadratic: $f(x) = 4x^2 - 2x + 6$ $= ax^2 + bx + c$.

In higher dimension, e.g. $x = (x_1, x_2, x_3)$:

- Linear: $3x_1 - 2x_2 + x_3 + 4$ $= ax_1 + bx_2 + cx_3 + d$ (sum of different terms; terms: constant or variables)
- Quadratic: $x_1^2 - 2x_1x_3 + 6x_2^2 + 7x_1 + 9$

$= ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + ex_2 + g$

(sum of terms; terms - constant / a linear term / square of a variable / pairwise of 2 variables)

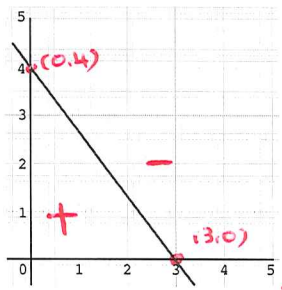
Linear functions and dot products

- SVM or Logistic Regression.

Based on linear decision boundaries.

i $\mathbb{R}^2$ **Linear separator** $4x_1 + 3x_2 = 12$:

(3,0) / (0,4) all satisfy the equation.



- suppose data in 2 dimensional space;
- one side of the boundary is plus, another is minus
- this particular line follows the equation $y = ax_1 + bx_2$

ii $\mathbb{R}^d$ For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, linear separators are of the form:

$$w_1x_1 + w_2x_2 + \cdots + w_dx_d = c.$$

Can write as $w \cdot x = c$, for $w = (w_1, \ldots, w_d)$.

∴ Dot Product allows us to express linear boundaries.

More general linear functions

- a function that takes a 4 dimensional vector $x_1 \sim x_4$, returns a linear function of these 4 features

A linear function from $\mathbb{R}^4$ to $\mathbb{R}$: $f(x_1, x_2, x_3, x_4) = 3x_1 - 2x_3$

$f(x_1, x_2, x_3, x_4) = (3, 0, -2, 0) \cdot (x_1, x_2, x_3, x_4)$

↳ write down a linear function using a dot product.

A linear function from $\mathbb{R}^4$ to $\mathbb{R}^3$: $f(x_1, x_2, x_3, x_4) = (4x_1 - x_2, x_3, -x_1 + 6x_4)$

Input. R^4. 4dimensional vectors.

Output. 3 numbers R^3.

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \to \begin{pmatrix} (4,-1,0,0)\cdot x \\ (0,0,1,0)\cdot x \\ (-1,0,0,6)\cdot x \end{pmatrix} = \begin{bmatrix} 4 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 6 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

Matrix × vector

The identity matrix

- the simplest possible linear function, is just sending x to itself ~ identity function.

The $d \times d$ **identity matrix** I_d sends each $x \in \mathbb{R}^d$ to itself.

$$I_d = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

$I_x = x$

Matrix-vector product

Product of matrix $M \in \mathbb{R}^{r \times d}$ and vector $x \in \mathbb{R}^d$:

matrix M

$$\begin{bmatrix} - M_1 - \\ - M_2 - \\ \vdots \\ - M_r - \end{bmatrix} \begin{bmatrix} | \\ x \\ | \end{bmatrix} = \begin{bmatrix} M_1 \cdot x \\ M_2 \cdot x \\ \vdots \\ M_r \cdot x \end{bmatrix}$$

$\leftarrow d \rightarrow$

$r \times d$ matrix $d \times 1$ vector $r \times 1$ matrix

∴ A matrix-vector product captures a linear function.

Matrix-matrix product

Product of matrix $A \in \mathbb{R}^{r \times k}$ and matrix $B \in \mathbb{R}^{k \times p}$:

$$\begin{bmatrix} - A_1 - \\ \vdots \\ - A_r - \end{bmatrix} \begin{bmatrix} | & & | \\ B^{(1)} & \cdots & B^{(p)} \\ | & & | \end{bmatrix} = \begin{bmatrix} & & \\ & \boxed{A_i B^{(j)}} & \\ & & \end{bmatrix}$$

A B AB j i

$\leftarrow k \rightarrow$ $\leftarrow p \rightarrow$

$r \times k$ $k \times p$ $r \times p$

the inner dimensions have to agree (k)

the size of product is given by the outer dimensions (r / p)

entry ij ~ entry row number i, col ~ j

ij entry of AB is simply a dot product of the ith row of A and j^{th} col of B

$$\begin{array}{c} \\ i \rightarrow \end{array}\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}\begin{array}{c} j \\ \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ -1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} -2 & 7 \\ -2 & 16 \end{bmatrix}$$

$2\times 3 \longleftrightarrow 3\times 2$ $\qquad$ $2\times 2.$

inner dimension

$(1,2,3)\cdot(1,0,-1) = -2$

$(1,2,3)\cdot(0,2,1) = 7$

$(4,5,6)\cdot(1,0,-1) = -2$

$(4,5,6)\cdot(0,2,-1) = 16$

Some special cases

For vector $x \in \mathbb{R}^d$, what are $x^T x$ and xx^T?

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

$$x^T x = x \cdot x = \|x\|^2$$

$$xx^T = \underset{d\times 1}{\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}} \cdot \underset{1\times d}{(x_1 \cdots x_d)} = \underset{d\times d}{(x_i x_j)}$$

Matrix products

If $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{k\times p}$, then AB is an $r \times p$ matrix with (i,j) entry

Dimension. $\quad$ i^{th} row of A $\quad$ j^{th} col of B

$$(AB)_{ij} = (\text{dot product of } i\text{th row of } A \text{ and } j\text{th column of } B) = \sum_{\ell=1}^{k} A_{i\ell} B_{\ell j}$$

- $I_k B = B$ and $A I_k = A$ → when multiply sth by identity matrix, it doesn't change.
- Can check: $(AB)^T = B^T A^T$
- For two vectors $u, v \in \mathbb{R}^d$, what is $u^T v$?

$u^T v$

$$\begin{pmatrix} u_1 \\ \vdots \\ u_d \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_d \end{pmatrix} = (u_1 \cdots u_d) \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_d \end{pmatrix}$$

$= u \cdot v$

$= u^T v$

$(AB)^T$ dimension $p\times r$

B^T $\quad p\times k$

A^T $\quad k\times r$

Associative but not commutative

- Multiplying matrices is **not commutative**: in general, $AB \neq BA$ → cannot change order

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

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$$

- But it is **associative**: $ABCD = (AB)(CD) = (A(BC))D$, etc. → can parenthesize the expression.

Example: if $x \in \mathbb{R}^d$ has length 2, what is $\underbrace{x^T x}_{\|x\|^2} x^T x x^T x$?

$= (\|x\|^2)^4 = \|x\|^{16}$

week 3-3 20 June 2018.

Linear algebra III
Square matrices as quadratic functions

- If the matrix is square; it has the same number of rows and cols, then it ~~requires~~ represents a quadratic function.

Topics we'll cover

1. Square matrices as quadratic functions
2. Special cases of square matrices: symmetric and diagonal
3. Determinant
4. Inverse

A special case

Recall: For vector $x \in \mathbb{R}^d$, we have $x^T x = \|x\|^2$. $= x^T \underbrace{I}_{} x$

What about $x^T M x$, for arbitrary $d \times d$ matrix M?

$$(x_1 \cdots x_d) \begin{bmatrix} & M & \end{bmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

$$1 \times d \text{ — } d \times d \text{ — } d \times 1$$

$$= \sum_{i,j} M_{ij} x_i x_j$$

What is $x^T M x$ for $M = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$?

2×2 ~ vectors x has to be 2-dimensional

$$(x_1 \; x_2) \underbrace{\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}}_{\substack{2\times 2 \\ \text{matrix}}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{\substack{2\times 1 \\ \text{vector}}} = \underbrace{(x_1 \; x_2)}_{1\times 2\ M} \underbrace{\begin{pmatrix} x_1 + 2x_2 \\ 3x_2 \end{pmatrix}}_{2\times 1\ M} = \underbrace{x_1^2 + 2x_1x_2 + 3x_2^2}_{1\times 1 \text{ number}}$$

$$\sum_{i,j} M_{ij} x_i x_j = M_{11} x_1 x_1 + M_{12} x_1 x_2 + M_{21} x_2 x_1 + M_{22} x_2 x_2$$

$$= 1 \cdot x_1^2 + 2 \cdot x_1 \cdot x_2 + 0 + 3 \cdot x_2^2$$

2-dimensional,

$$= x_1^2 + 2x_1x_2 + 3x_2^2$$

so sum over one and two.

Quadratic functions

Let M be any $d \times d$ (**square**) matrix.
For $x \in \mathbb{R}^d$, the mapping $x \mapsto x^T M x$ is a **quadratic function** from $\mathbb{R}^d$ to $\mathbb{R}$:

$$\underset{\mathbb{R}^d}{x} \longrightarrow x^T M x = \sum_{i,j=1}^{d} M_{ij} x_i x_j.$$

What is the quadratic function associated with $M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 3 & 4 & 5 \end{pmatrix}$?

M_{11} M_{22} M_{31}

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \rightarrow 1 \cdot x_1 x_1 + 2 x_2 x_2 + 3 x_3 x_1 + 4 x_3 x_2 + 5 x_3 x_3$$

$$= x_1^2 + 2x_2^2 + 3x_1x_3 + 4x_2x_3 + 5x_3^2$$

∴ a quadratic function associated with this matrix

Reverse ~ have a quadratic function that takes a 2-dimensional vector.
Write the quadratic function $f(x_1, x_2) = 1x_1^2 + 2x_1x_2 + 3x_2^2$ using matrices and vectors.

~ how do we write this fun using a matrix?

2-dimensional vector - 2x2 matrix

$$\begin{bmatrix} 1 & a \\ b & 3 \end{bmatrix} \quad a+b=2$$

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

all possibilities

Special cases of square matrices

- matrix remains the same if you reflect it in diagonal

- **Symmetric**: $M = M^T$

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 4 \\ 3 & 4 & 6 \end{pmatrix}$$

non-symmetrix

- **Diagonal**: $M = \text{diag}(m_1, m_2, \ldots, m_d)$
  - 0 everywhere except along the diagonal over there.

$$\text{diag}(1, 4, 7) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 7 \end{pmatrix}$$

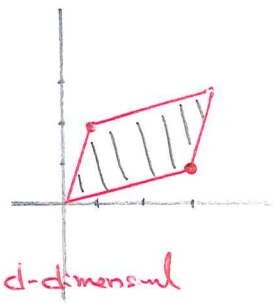
- one of the key quantities associated with square matrix ~ determinant

Determinant of a square matrix

Determinant of $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $|A| = ad - bc$.

Example: $A = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$

$|A| = 3 \times 2 - 1 \times 1$
$= 5$



- take two points (3,1) (1,2)
- extend them to set a parallelogram
- area of this parallelogram is 5.

- also works for d-dimensional matrix
- d×d matrix, look at the rows, plot this points in a d-dimensional space
- extend them into a parallelogram
- the volume of objects is the determinant of matrix

- key quantity of square matrix ~ inverse matrix

Inverse of a square matrix

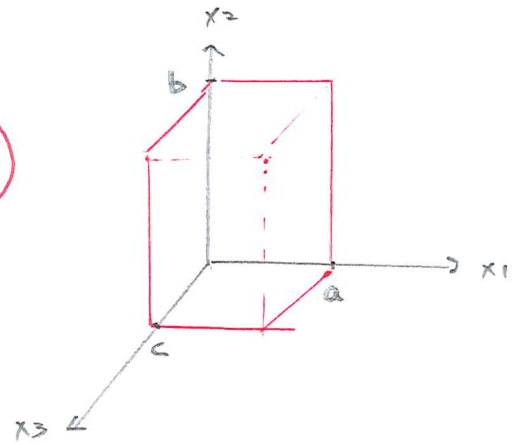
The **inverse** of a $d \times d$ matrix A is a $d \times d$ matrix B for which $AB = BA = I_d$.
Notation: A^{-1}.

- B has to also be a $d \times d$ matrix

Example: if $A = \begin{pmatrix} 1 & 2 \\ -2 & 0 \end{pmatrix}$ then $A^{-1} = \begin{pmatrix} 0 & -1/2 \\ 1/2 & 1/4 \end{pmatrix}$. Check!

$$\begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1/2 \\ 1/2 & 1/4 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{I_d}$$

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$



Volume of this is abc

, which is the determinant of matrix

$|A| = abc - 0 \cdot b \cdot 0$

$= abc$.

The determinant of diagonal matrix

= Product of diagonal element

Inverse of a square matrix, cont'd

The **inverse** of a $d \times d$ matrix A is a $d \times d$ matrix B for which $AB = BA = I_d$.
Notation: A^{-1}.

singular = not invertible.

- Not all square matrices have an inverse
- Square matrix A is invertible if and only if $|A| \neq 0$ $\quad |A| = 0$ ~ singular.
- What is the inverse of $A = \text{diag}(a_1, \ldots, a_d)$?

$$\begin{pmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_d \end{pmatrix}$$

$d \times d$

determinant $= a_1 \cdot a_2 \cdots a_d$

Need $a_i \neq 0$

$$\begin{pmatrix} 1/a_1 & & 0 \\ & \ddots & \\ 0 & & 1/a_d \end{pmatrix}$$

inverse matrix

A square matrix that is not invertible is called singular / degenerate.

A square matrix is singular only if $|A| = 0$.

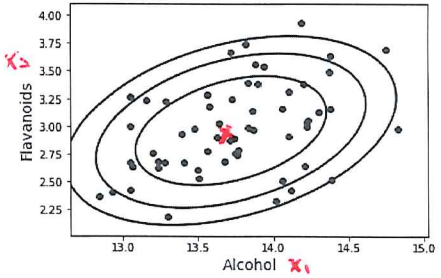
week 3-4 25/ June /2018.

The multivariate Gaussian

Topics we'll cover

1. Functional form of the density
2. Special case: diagonal Gaussian
3. Special case: spherical Gaussian
4. Fitting a Gaussian to data

Recall: the bivariate Gaussian



- real data - actual data points
- fit a Gaussian to them, comprising μ of each features and variance betn x_1 and x_2.

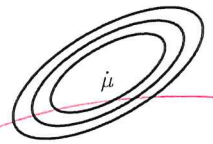
Bivariate Gaussian, parametrized by:

$$\text{mean } \mu = \begin{pmatrix} 13.7 \\ 3.0 \end{pmatrix} \text{ and covariance matrix } \Sigma = \begin{pmatrix} 0.20 & 0.06 \\ 0.06 & 0.12 \end{pmatrix}$$

var(x_1) cov(x_1, x_2)
var(x_2)

- μ, the center of the point, the highest point of density
- the counter lines ellipses of the density
- upwards - positive correlation between 2 features

The multivariate Gaussian



$N(\mu, \Sigma)$: Gaussian in $\mathbb{R}^d$ - d-dimensional space.

- mean: $\mu \in \mathbb{R}^d$
- covariance: $d \times d$ matrix Σ

Generates points $X = (X_1, X_2, \ldots, X_d)$. - d features

$$\Sigma = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) & \text{cov}(x_1, x_d) \\ & \ddots & \\ & & \text{var}(x_d) \end{bmatrix}$$

d×d matrix

- μ is the vector of coordinate-wise means:

$$\mu_1 = \mathbb{E}X_1, \ \mu_2 = \mathbb{E}X_2, \ldots, \ \mu_d = \mathbb{E}X_d.$$

- Σ is a matrix containing all pairwise covariances:

$$\Sigma_{ij} = \Sigma_{ji} = \text{cov}(X_i, X_j) \quad \text{if } i \neq j$$
$$\Sigma_{ii} = \text{var}(X_i)$$

- the i,j entry of this matrix is the covariance between x_i and x_j

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

- normalize zone
- a constant

- only part depends on x
- when $\mu = 0$
  $\exp(-\frac{1}{2}(x)^T \cdot M(x))$ ~ quadratic function
- so density depends on quadratic func of x.

Special case: independent features

Suppose the X_i are independent, and $\text{var}(X_i) = \sigma_i^2$. $\text{Cov}(X_i, X_j) = 0$

What is the covariance matrix Σ, and what is its inverse Σ^{-1}?

$$\Sigma = \begin{bmatrix} \text{Var}(X_1) & & 0 \\ & \ddots & \\ 0 & & \text{Var}(X_d) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix} = \text{diag}(\sigma_1^2 \cdots \sigma_d^2)$$

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_d^2 \end{bmatrix}$$

Diagonal Gaussian

only need $d+d = 2d$ parameters $\sim d^2$

Diagonal Gaussian: the X_i are independent, with variances σ_i^2. Thus

$$\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2) \quad \text{(off-diagonal elements zero)}$$

Each X_i is an independent one-dimensional Gaussian $N(\mu_i, \sigma_i^2)$:

times ... point x's

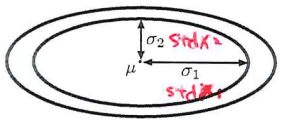
each feature x_i fits a Gaussian of one-dimensional.

$$\Pr(x) = \Pr(x_1)\Pr(x_2)\cdots\Pr(x_d) = \frac{1}{(2\pi)^{d/2}\sigma_1\cdots\sigma_d}\exp\left(-\sum_{i=1}^{d}\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right)$$

one-dimensional density of point x_i

density of d-dimensional point x

Contours of equal density are **axis-aligned ellipsoids** centered at μ:



$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \sigma_1^2 > \sigma_2^2$$

- Covariance Matrix = Diagonal Matrix

  correlation = 0

- the ellipse is axis aligned, parallel to the coordinate directions.

Even more special case: spherical Gaussian

The X_i are independent and all have the same variance σ^2.

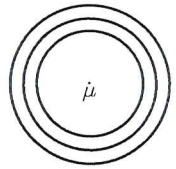
$$\Sigma = \sigma^2 I_d = \text{diag}(\sigma^2, \sigma^2, \ldots, \sigma^2) \quad \text{(diagonal elements } \sigma^2\text{, rest zero)}$$

Each X_i is an independent univariate Gaussian $N(\mu_i, \sigma^2)$:

$$\Pr(x) = \Pr(x_1)\Pr(x_2)\cdots\Pr(x_d) = \frac{1}{(2\pi)^{d/2}\sigma^d}\exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right)$$

- the density of point x depends only on the distance from x to μ.
- the points with equal density are the points that lies on a sphere.

Density at a point depends only on its distance from μ:



How to fit a Gaussian to data

Fit a Gaussian to data points $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^d$.

- Empirical mean

$$\mu = \frac{1}{m}\left(x^{(1)} + \cdots + x^{(m)}\right)$$

- Empirical covariance matrix has i, j entry:

$$\Sigma_{ij} = \left(\frac{1}{m}\sum_{k=1}^{m} x_i^{(k)} x_j^{(k)}\right) - \mu_i\mu_j$$

$$\text{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)$$

- the only thing need to compute is the average value of x_i ~~and~~ x_j in the data set.

week 3-5 26 June 2015.

Gaussian generative models

- Gaussian ~ probability distribution
- using these distribution to build classifers.

Topics we'll cover

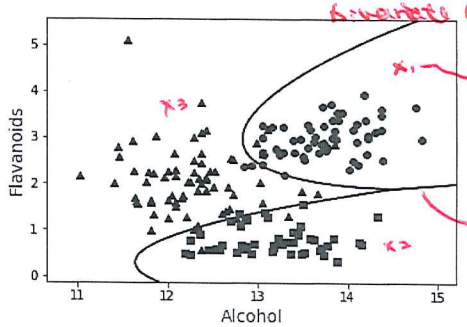
1. Classification using multivariate Gaussian generative modeling
2. The form of the decision boundaries

Back to the winery data

one dimensional Gaussian

Go from 1 to 2 features: test error goes from 29% to 8%.

bivariate Gaussian



x_1 → classified as '0'

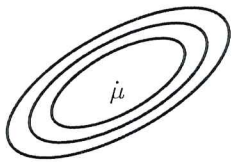
→ classification boundary

With all 13 features: test error rate goes to zero. → better classifier

↳ fit a multivariant Gaussian to 13 features for winery one,
" " " two,
" " " three.

- Then apply Bayes Rule classification.

The multivariate Gaussian



$N(\mu, \Sigma)$: Gaussian in $\mathbb{R}^d$

- mean: $\mu \in \mathbb{R}^d$
- covariance: $d \times d$ matrix Σ

- Bring down the algorithm.

$\log P(r) = \text{constant} - \frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu))$

↑ quadratic.

$$\text{Density} \quad p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(\underbrace{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}_{\text{Quadratic function}}\right)$$

If we write $S = \Sigma^{-1}$ then S is a $d \times d$ matrix and

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \sum_{i,j} S_{ij}(x_i-\mu_i)(x_j-\mu_j),$$

a **quadratic function** of x.

↳ if use Gaussian distributions in the generative approach to classification, the decision boundaries are generally quadratic.

Binary classification with Gaussian generative model

- Estimate class probabilities π_1, π_2 $\pi_1 = 0.6$ $\pi_2 = 0.4$ - 60% of data from class 1.
- Fit a Gaussian to each class: $P_1 = N(\mu_1, \Sigma_1)$, $P_2 = N(\mu_2, \Sigma_2)$

Given a new point x, predict class 1 if

$$\pi_1 P_1(x) \geq \pi_2 P_2(x) \quad \Leftrightarrow \quad \underbrace{x^T M x + 2 w^T x \geq \theta}_{\text{decision rules.}},$$

threshold theta. i.e. 0.01-0.6

where:

$$M = \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1})$$
$$w = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2$$

and θ is a threshold depending on the various parameters.

Linear or **quadratic** decision boundary.

↳ decision boundary is separating region between two zones; which corresponds to the equation $x^T M x + 2 w^T x = 0$

↳ Sometimes, when $M = 0$, the covariance matrix are identical. then $2 w^T x$, just a linear boundary.

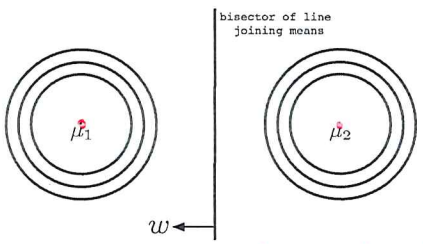
Linear Boundary

Common covariance: $\Sigma_1 = \Sigma_2 = \Sigma$

Linear decision boundary: choose class 1 if

$$x \cdot \underbrace{\Sigma^{-1}(\mu_1 - \mu_2)}_{w} \geq \theta.$$

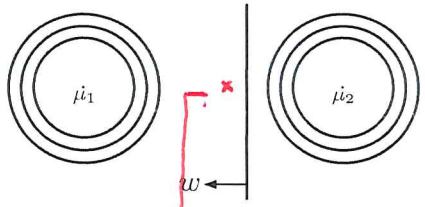
Example 1: Spherical Gaussians with $\Sigma = I_d$ and $\pi_1 = \pi_2$. $\Sigma_1 = \Sigma_2$ (identical matrix).



↳ decision boundary.

Perpendicular bisector between mid pts.

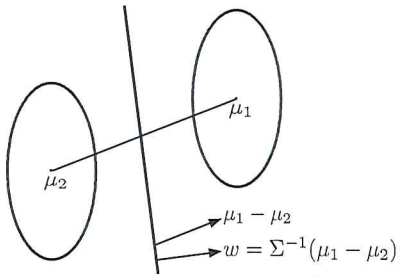
Example 2: Again spherical, but now $\pi_1 > \pi_2$. $\Sigma_1 = \Sigma_2 = \Sigma$.



↳ decision boundary moves a bit right.

↳ classified as class 1; because class 1 occurs more frequently than class 1 ($\pi_1 > \pi_2$)

Example 3: Non-spherical. $\Sigma_1 = \Sigma_2 = \Sigma$. boundary ~ linear.



↳ decision boundary. skewed slightly

∴ Classification rule: $w \cdot x \geq \theta$ (linear)

- Choose w as above
- Common practice: fit θ to minimize training or validation error

Quadratic boundary

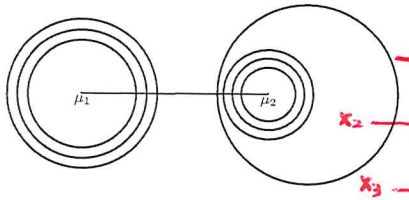
Different covariances: $\Sigma_1 \neq \Sigma_2$

Quadratic boundary: choose class 1 if $x^T M x + 2w^T x \geq \theta$, where:

$$M = \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1})$$
$$w = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2$$

Example 1: $\Sigma_1 = \sigma_1^2 I_d$ and $\Sigma_2 = \sigma_2^2 I_d$ with $\sigma_1 > \sigma_2$

classified as class 1. ← x_1



→ decession boundary is sphere.

x_2 → classified as Class 0.

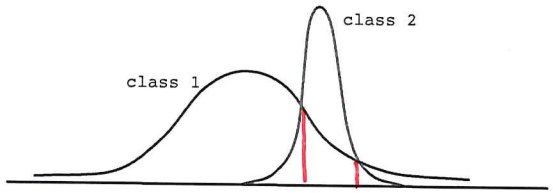
x_3 → classified as class 1

Larger variance

4I matrix

I matrix

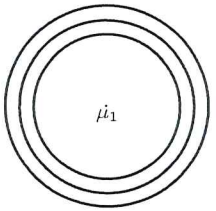
Example 2: Same thing in 1-d. $\mathcal{X} = \mathbb{R}$.



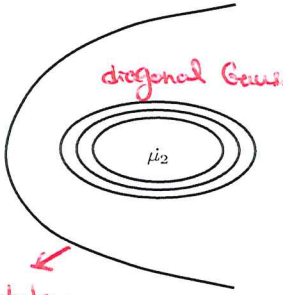
classified as Class 1 | classified as 0 | class 1

Example 3: A parabolic boundary.

spherical Gaussian



diagonal Gaussian



parabola.

Multiclass discriminant analysis

k classes: weights π_j, class-conditional densities $P_j = N(\mu_j, \Sigma_j)$.

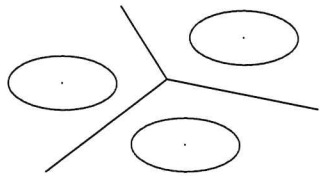
Each class has an associated **quadratic** function

$$f_j(x) = \log(\pi_j P_j(x))$$

To classify point x, pick $\arg\max_j f_j(x)$.

If $\Sigma_1 = \cdots = \Sigma_k$, the boundaries are **linear**.

- when a new point comes.
- each class evaluate its quadratic fnc, get a number.
- one with the largest number is the winner.



Week 3-6. 26/June/2018.

More generative modeling

- looking at a variety of one-dimensional distributions.
- see how combine these to deal with higher dimensional data.

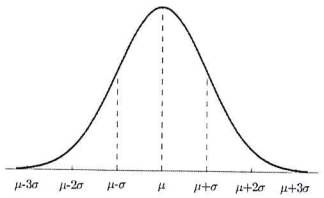
Topics we'll cover

1. Beyond Gaussians
2. A variety of univariate distributions
3. Moving to higher dimension

Classification with generative models

- Fit a **distribution** to each class separately
- Use Bayes' rule to classify new data

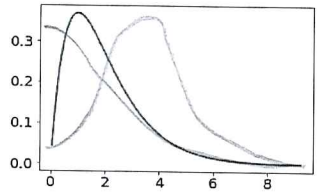
What distribution to use? Are Gaussians enough?



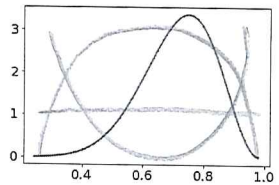
- when Gaussian distribution would not be suitable.

Exponential families of distributions

- chi-square dis.
- exponential dis (special cases of GAMMA dis)
- when data is real valued numeric but always positive.

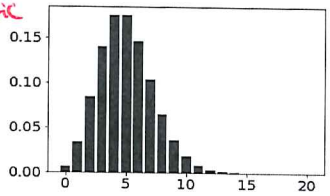


GAMMA



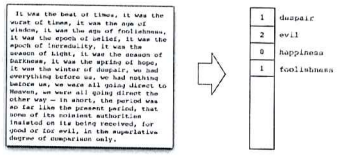
BETA

- when data is real valued but lies in a specific interval.



POISSON

- when data is all integers. i.e. counts (positive always).
- single parameter (by mean of the distribution)



CATEGORICAL

- data is non-numeric.
- assign probability of each outcomes.

- Combine these distributions.
- Deal with higher dimensional data.

Multivariate distributions

We've described a variety of distributions for **one-dimensional** data.
What about higher dimensions?

1. **Naive Bayes**: Treat coordinates as independent.
   For $x = (x_1, \ldots, x_d)$, fit separate models $\Pr_i$ to each x_i, and assume

   $$\Pr(x_1, \ldots, x_d) = \Pr_1(x_1)\Pr_2(x_2)\cdots\Pr_d(x_d).$$

   This assumption is typically inaccurate.
2. **Multivariate Gaussian**.
   Model correlations between features: we've seen this in detail.
3. **Graphical models**.
   Arbitrary dependencies between coordinates.