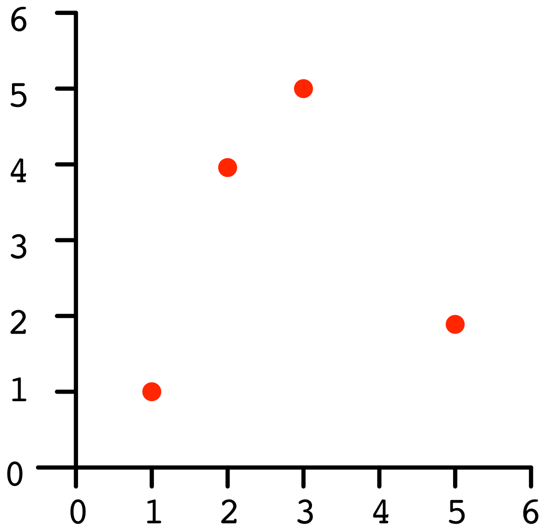
Linear algebra I
Basic vector-matrix notation, and dot products

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



Matrix-vector notation

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

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

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

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

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

Transpose of vectors and matrices

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

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

- $(A^T)_{ij} = A_{ji}$
- $(A^T)^T = A$

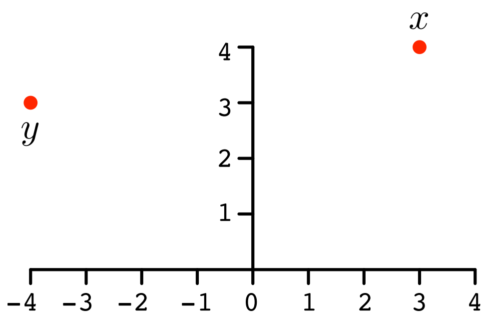
Adding and subtracting vectors and matrices

Dot product of two vectors

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

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

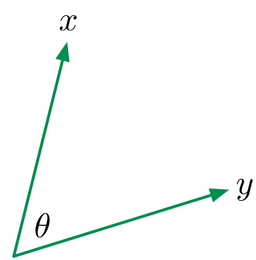
What is the dot product between these two vectors?



Dot products and angles

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:



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

x is **orthogonal** (at right angles) to y if and only if $x \cdot y = 0$
When x, y are **unit vectors** (length 1): $\cos\theta = x \cdot y$
What is $x \cdot x$?

Linear algebra II
Linear functions and matrix products

Topics we'll cover

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

Linear and quadratic functions

In one dimension:

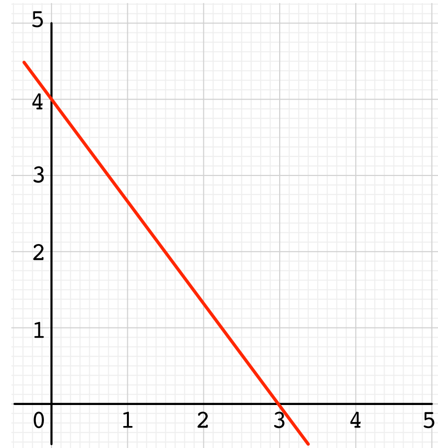
- Linear: $f(x) = 3x + 2$
- Quadratic: $f(x) = 4x^2 - 2x + 6$

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

- Linear: $3x_1 - 2x_2 + x_3 + 4$
- Quadratic: $x_1^2 - 2x_1x_3 + 6x_2^2 + 7x_1 + 9$

Linear functions and dot products

Linear separator $4x_1 + 3x_2 = 12$:



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

$$w_1 x_1 + w_2 x_2 + \cdots + w_d x_d = c.$$

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

More general linear functions

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

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

Matrix-vector product

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

The identity matrix

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

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

Matrix-matrix product

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

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

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

- $I_k B = B$ and $A \, I_k = A$
- Can check: $(AB)^T = B^T A^T$
- For two vectors $u, v \in \mathbb{R}^d$, what is $u^T v$?

Some special cases

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

Associative but not commutative

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

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

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

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

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

Linear algebra III
Square matrices as quadratic functions

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

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

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

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 $\mathbb{R}$:

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

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

Write the quadratic function $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^2$ using matrices and vectors.

Special cases of square matrices

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

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

- **Diagonal**: $M = \mathrm{diag}(m_1, m_2, \ldots, m_d)$

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

Determinant of a square matrix

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

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

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}.

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!

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}.

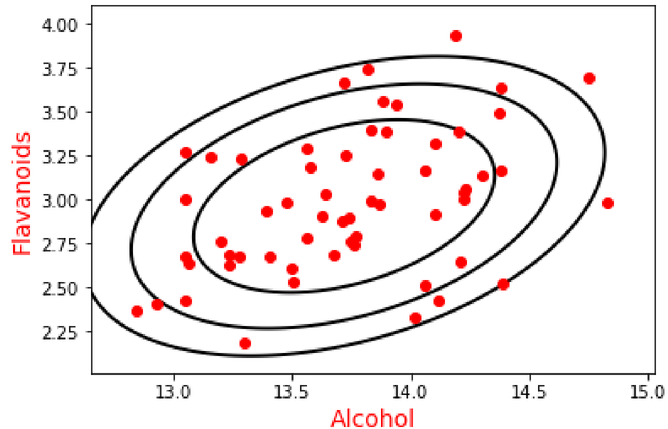
- Not all square matrices have an inverse
- Square matrix A is invertible if and only if $|A| \neq 0$
- What is the inverse of $A = \text{diag}(a_1, \ldots, a_d)$?

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

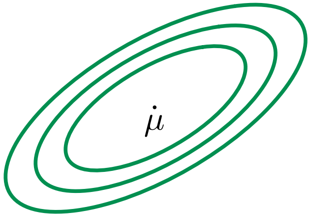
Recall: the bivariate Gaussian



Bivariate Gaussian, parametrized by:

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

The multivariate Gaussian



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

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

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

- 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:

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

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

Special case: independent features

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

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

Diagonal Gaussian

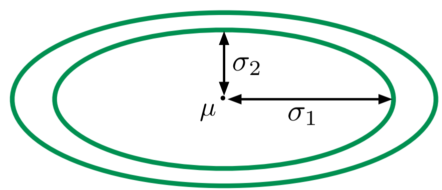
Diagonal Gaussian: the X_i are independent, with variances σ_i^2. Thus

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

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

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

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



Even more special case: spherical Gaussian

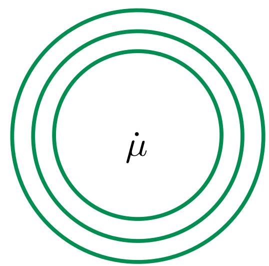
The X_i are independent and all have the same variance σ^2.

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

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

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

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:

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

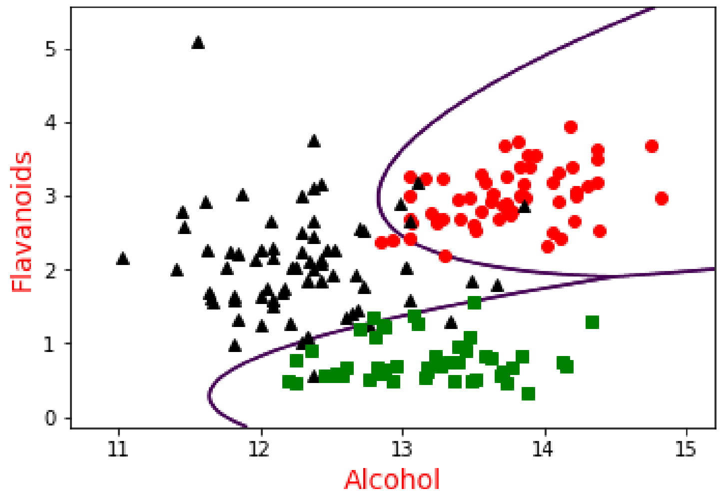
Gaussian generative models

Topics we'll cover

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

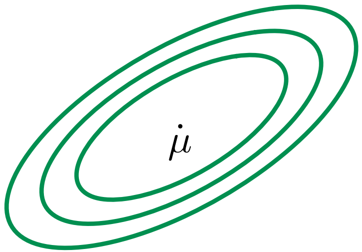
Back to the winery data

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



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

The multivariate Gaussian



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

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

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

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.

Binary classification with Gaussian generative model

- Estimate class probabilities π_1, π_2
- 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

$$\pi_1 P_1(x) > \pi_2 P_2(x) \quad \Leftrightarrow \quad x^T M x + 2 w^T x \geq \theta,$$

where:

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

and θ is a threshold depending on the various parameters.

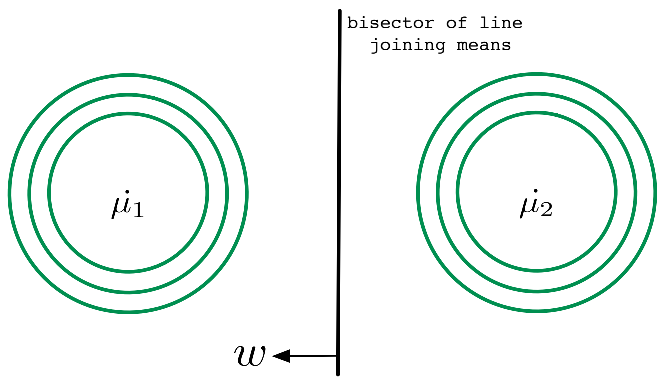
Linear or **quadratic** decision boundary.

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

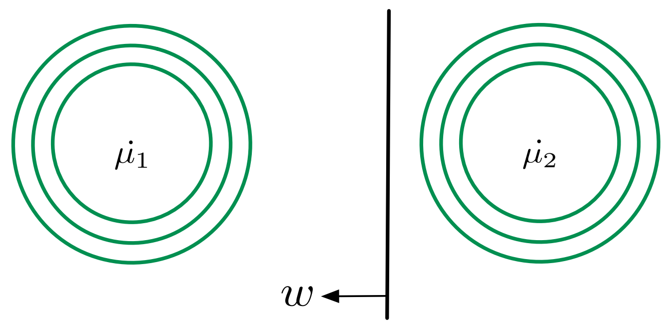
Linear decision boundary: choose class 1 if

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

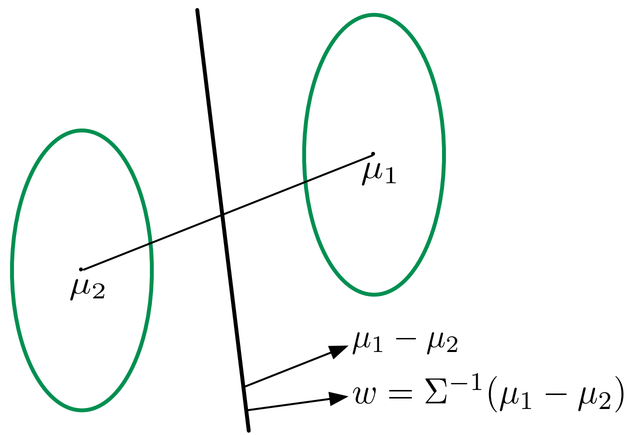
Example 1: Spherical Gaussians with $\Sigma = I_d$ and $\pi_1 = \pi_2$.



Example 2: Again spherical, but now $\pi_1 > \pi_2$.



Example 3: Non-spherical.



Classification rule: $w \cdot x \geq \theta$

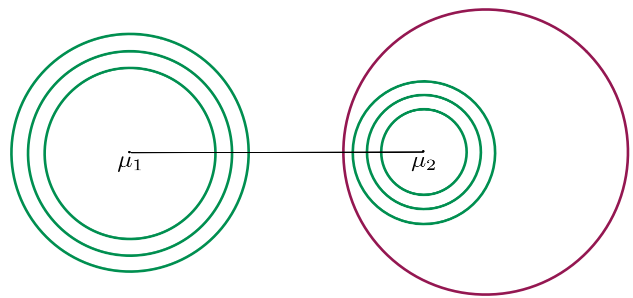
- Choose w as above
- Common practice: fit θ to minimize training or validation error

Different covariances: $\Sigma_1 \neq \Sigma_2$

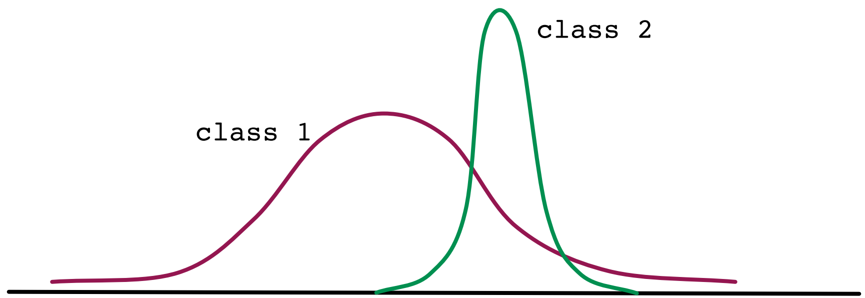
Quadratic boundary: choose class 1 if $x^T M x + 2w^T x \geq \theta$, where:

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

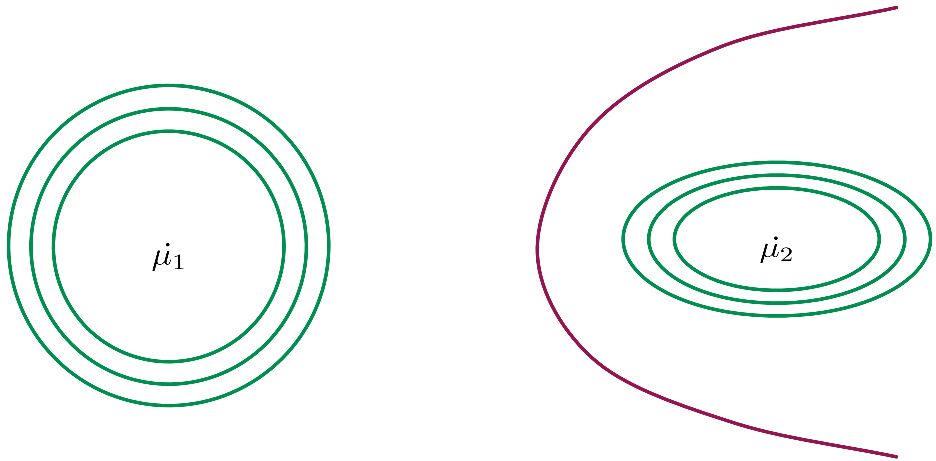
Example 1: $\Sigma_1 = \sigma_1^2 I_d$ and $\Sigma_2 = \sigma_2^2 I_d$ with $\sigma_1 > \sigma_2$



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



Example 3: A parabolic boundary.



Multiclass discriminant analysis

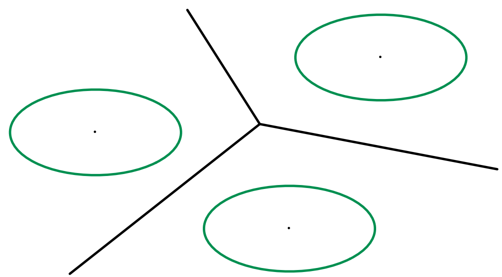
k classes: weights π_j, class-conditional densities $P_j = N(\mu_j, \Sigma_j)$.

Each class has an associated **quadratic** function

$$f_j(x) = \log\left(\pi_j \, P_j(x)\right)$$

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

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



More generative modeling

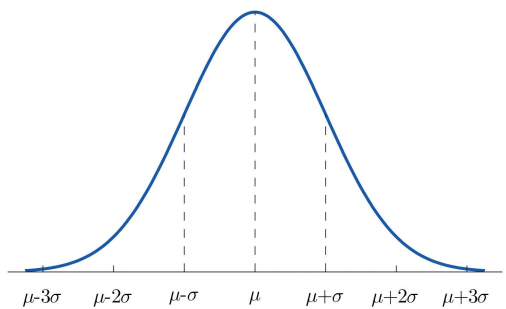
Topics we'll cover

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

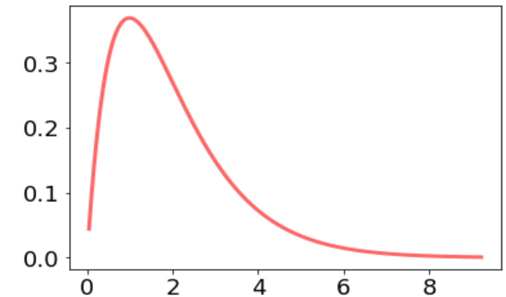
Classification with generative models

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

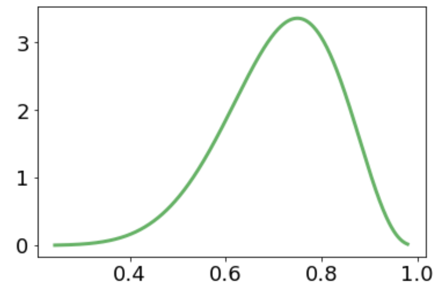
What distribution to use? Are Gaussians enough?



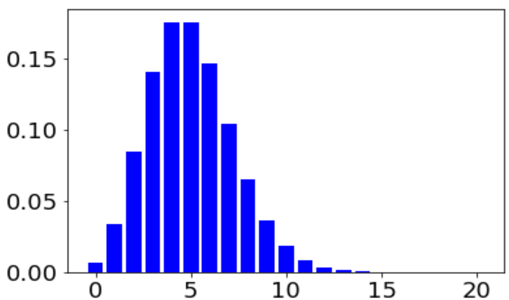
Exponential families of distributions



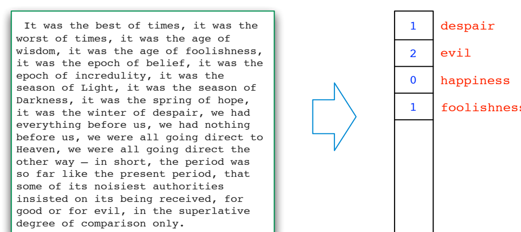
GAMMA



BETA



POISSON



CATEGORICAL

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.