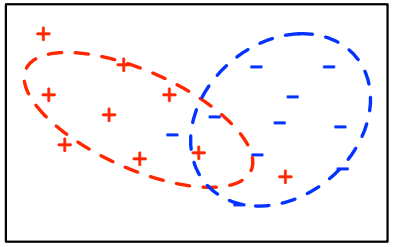
The generative approach to classification

CSE 250B

The generative approach to classification



The learning process:

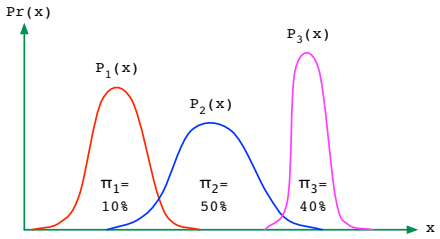
- Fit a probability distribution to each class, individually

To classify a new point:

- Which of these distributions was it most likely to have come from?

Generative models

Example:
Data space $\mathcal{X} = \mathbb{R}$
Classes/labels $\mathcal{Y} = \{1, 2, 3\}$



For each class j, we have:

- the probability of that class, $\pi_j = \Pr(y = j)$
- the distribution of data in that class, $P_j(x)$

Overall **joint distribution**: $\Pr(x, y) = \Pr(y)\Pr(x|y) = \pi_y P_y(x)$.

To classify a new x: pick the label y with largest $\Pr(x, y)$

A classification problem

You have a bottle of wine whose label is missing.

Which winery is it from, 1, 2, or 3?

Solve this problem using visual and chemical features of the wine.

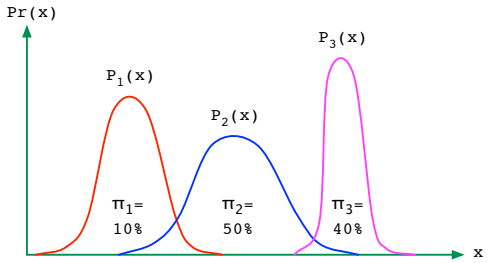
The data set

Training set obtained from 130 bottles

- Winery 1: 43 bottles
- Winery 2: 51 bottles
- Winery 3: 36 bottles
- For each bottle, 13 features:
  'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash', 'Magnesium',
  'Total phenols', 'Flavanoids', 'Nonflavanoid phenols',
  'Proanthocyanins',
  'Color intensity', 'Hue', 'OD280/OD315 of diluted wines',
  'Proline'

Also, a separate test set of 48 labeled points.

Recall: the generative approach



For any data point $x \in \mathcal{X}$ and any candidate label j,

$$\Pr(y = j|x) = \frac{\Pr(y = j)\Pr(x|y = j)}{\Pr(x)} = \frac{\pi_j P_j(x)}{\Pr(x)}$$

Optimal prediction: the class j with largest $\pi_j P_j(x)$.

Fitting a generative model

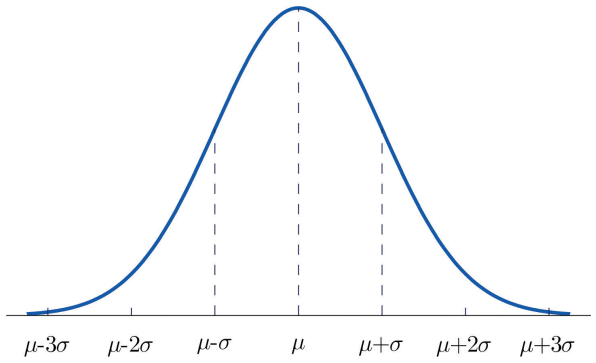
Training set of 130 bottles:

- Winery 1: 43 bottles, winery 2: 51 bottles, winery 3: 36 bottles
- For each bottle, 13 features: 'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash','Magnesium', 'Total phenols', 'Flavanoids', 'Nonflavanoid phenols', 'Proanthocyanins', 'Color intensity', 'Hue', 'OD280/OD315 of diluted wines', 'Proline'

Class weights:
$\pi_1 = 43/130 = 0.33, \quad \pi_2 = 51/130 = 0.39, \quad \pi_3 = 36/130 = 0.28$

Need distributions P_1, P_2, P_3, one per class.
Base these on a single feature: 'Alcohol'.

The univariate Gaussian

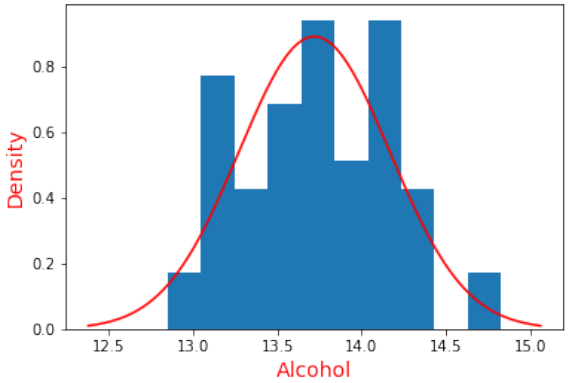


The Gaussian $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

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

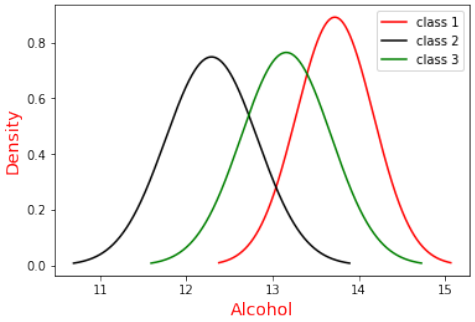
The distribution for winery 1

Single feature: 'Alcohol'



Mean $\mu = 13.72$, Standard deviation $\sigma = 0.44$ (variance 0.20)

All three wineries



- $\pi_1 = 0.33$, $P_1 = N(13.7, 0.20)$
- $\pi_2 = 0.39$, $P_2 = N(12.3, 0.28)$
- $\pi_3 = 0.28$, $P_3 = N(13.2, 0.27)$

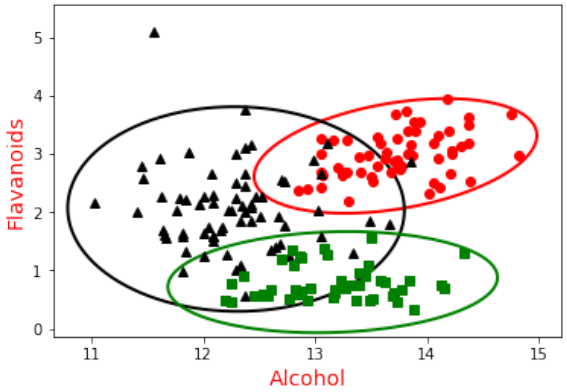
To classify x: Pick the j with highest $\pi_j P_j(x)$

Test error: 14/48 = 29%

What if we use **two** features?

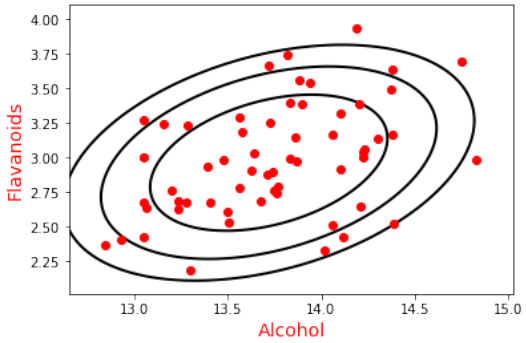
Why it helps to add features

Better **separation** between the classes!



Error rate drops from 29% to 8%.

The bivariate Gaussian



Model class 1 by a bivariate Gaussian, parametrized by:

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

Dependence between two random variables

Suppose X_1 has mean μ_1 and X_2 has mean μ_2.

Can measure dependence between them by their **covariance**:

- $\text{cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] = \mathbb{E}[X_1 X_2] - \mu_1 \mu_2$
- Maximized when $X_1 = X_2$, in which case it is $\text{var}(X_1)$.
- It is at most $\text{std}(X_1)\text{std}(X_2)$.

The bivariate (2-d) Gaussian

A distribution over $(x_1, x_2) \in \mathbb{R}^2$, parametrized by:

- **Mean** $(\mu_1, \mu_2) \in \mathbb{R}^2$, where $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_2)$
- **Covariance matrix** $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ where

$$\left\{ \begin{array}{c} \Sigma_{11} = \text{var}(X_1) \\ \Sigma_{22} = \text{var}(X_2) \\ \Sigma_{12} = \Sigma_{21} = \text{cov}(X_1, X_2) \end{array} \right\}$$

Density is highest at the mean,
falls off in ellipsoidal contours.

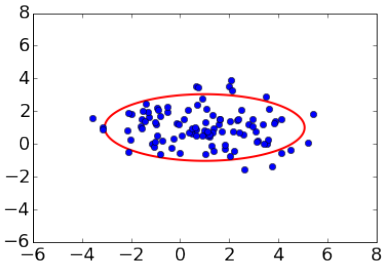
Density of the bivariate Gaussian

- **Mean** $(\mu_1, \mu_2) \in \mathbb{R}^2$, where $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_2)$
- **Covariance matrix** $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

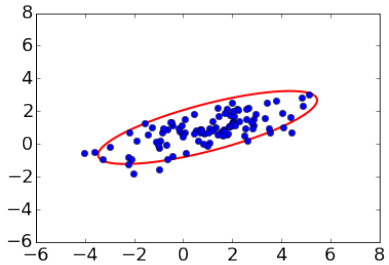
$$\text{Density} \quad p(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

Bivariate Gaussian: examples

In either case, the mean is $(1, 1)$.



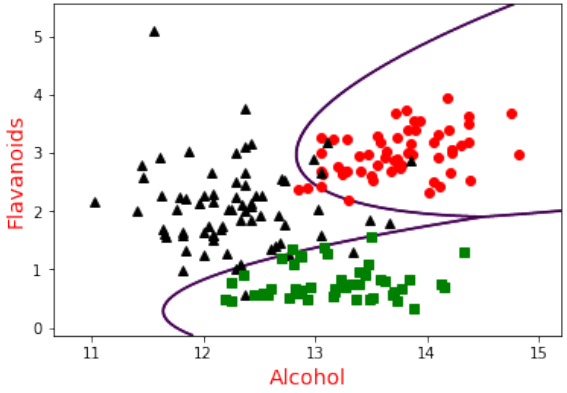
$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 4 & 1.5 \\ 1.5 & 1 \end{bmatrix}$$

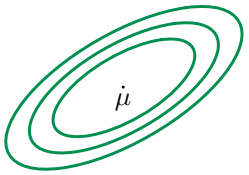
The decision boundary

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



What kind of function is this? And, can we use more features?

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

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

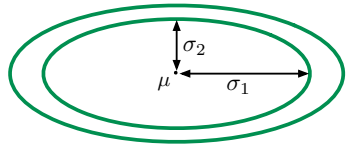
Diagonal Gaussian: the X_i are independent, with variances σ_i^2.

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

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

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

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



Even more special case: spherical Gaussian

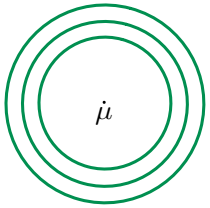
The X_i are independent and all have the same variance σ^2.

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

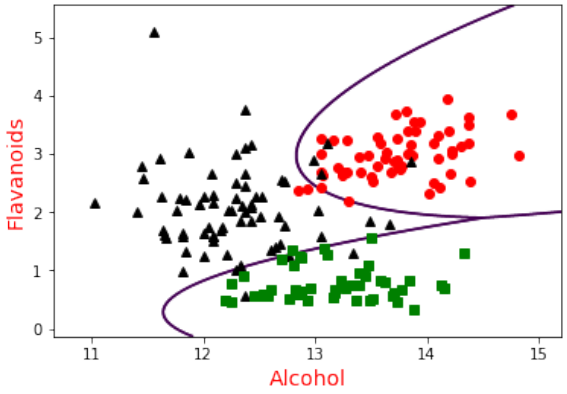
$$\mu = \frac{1}{m}\left(x^{(1)} + \cdots + x^{(m)}\right)$$

- Empirical covariance matrix has i, j entry:

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

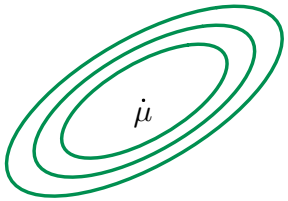
Back to the winery data

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

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

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

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

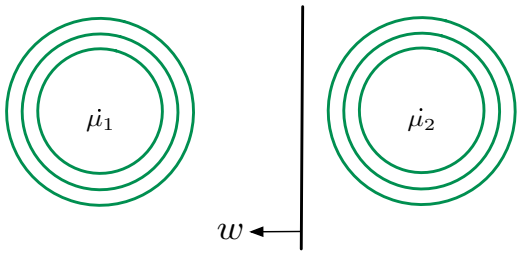
where:

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

and θ is a threshold depending on the various parameters.

Linear or **quadratic** decision boundary.

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

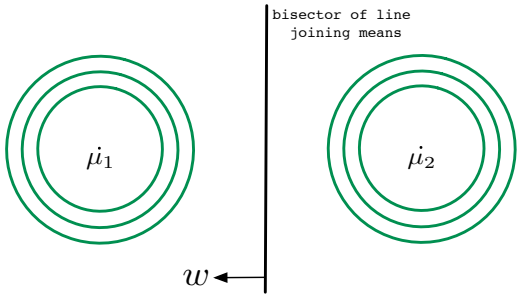


Common covariance: $\Sigma_1 = \Sigma_2 = \Sigma$

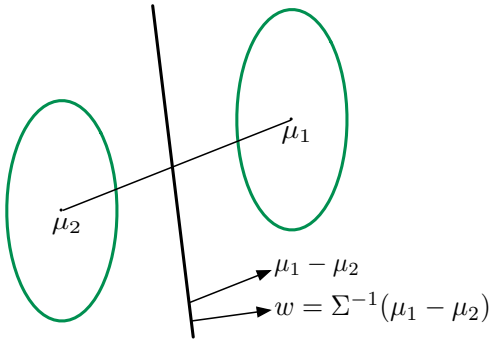
Linear decision boundary: choose class 1 if

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

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



Example 3: Non-spherical.



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

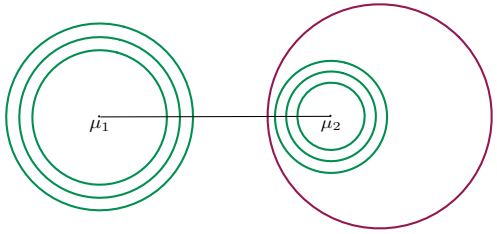
- Choose w as above
- Common practice: fit θ to minimize training or validation error

Different covariances: $\Sigma_1 \neq \Sigma_2$

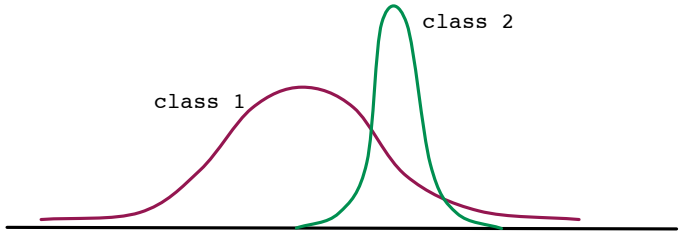
Quadratic boundary: choose class 1 if $x^T M x + 2w^T x \geq \theta$, where:

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

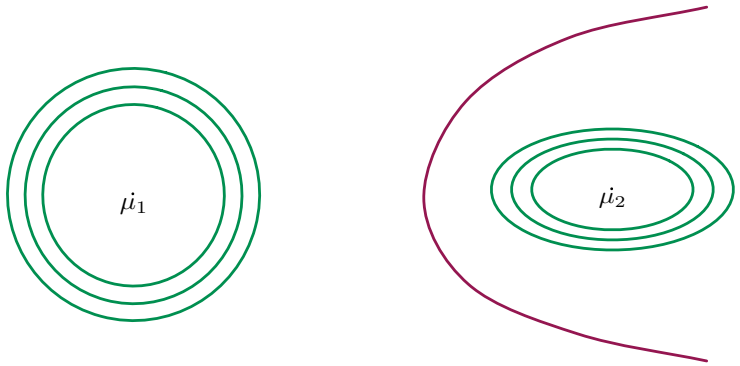
Example 1: $\Sigma_1 = \sigma_1^2 I_d$ and $\Sigma_2 = \sigma_2^2 I_d$ with $\sigma_1 > \sigma_2$



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

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

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

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