Multivariate generative modeling

The multivariate Gaussian

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

\[ N(\mu, \Sigma): \text{Gaussian in } \mathbb{R}^d \]
- mean: \( \mu \in \mathbb{R}^d \)
- covariance: \( d \times d \) matrix \( \Sigma \)

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

- \( \mu \) is the vector of coordinate-wise means:
  \[
  \mu_1 = \mathbb{E}X_1, \ \mu_2 = \mathbb{E}X_2, \ldots, \ \mu_d = \mathbb{E}X_d.
  \]

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

Special case: independent features

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

What is the covariance matrix \( \Sigma \), and what is its inverse \( \Sigma^{-1} \)?
Diagonal Gaussian

**Diagonal Gaussian:** the $X_i$ are independent, with variances $\sigma_i^2$. 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 $N(\mu_i, \sigma_i^2)$:

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

Contours of equal density are **axis-aligned ellipsoids** centered at $\mu$:

![Axis-aligned ellipsoids](image)

Even more special case: **spherical Gaussian**

The $X_i$ are independent and all have the same variance $\sigma^2$.

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

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

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

Density at a point depends only on its distance from $\mu$:

![Density ellipsoids](image)
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

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

\[ N(\mu, \Sigma): \text{Gaussian in } \mathbb{R}^d \]
- mean: \( \mu \in \mathbb{R}^d \)
- covariance: \( d \times d \) matrix \( \Sigma \)

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 $\pi_1, \pi_2$
- 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

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

where:

$$M = \frac{1}{2}(\Sigma^{-1}_2 - \Sigma^{-1}_1)$$
$$w = \Sigma^{-1}_1 \mu_1 - \Sigma^{-1}_2 \mu_2$$

and $\theta$ 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

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

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

![Diagram showing the bisector of a line joining the means of two spherical Gaussians with common covariance.](image-url)
Example 2: Again spherical, but now $\pi_1 > \pi_2$.

\[
\mu_1 \quad \mu_2
\]

$\mathbf{w}$

Example 3: Non-spherical.

\[
\mu_1 \quad \mu_2
\]

\[
\mu_1 - \mu_2
\]

\[
w = \Sigma^{-1}(\mu_1 - \mu_2)
\]

Classification rule: $\mathbf{w} \cdot \mathbf{x} \geq \theta$

- Choose $\mathbf{w}$ as above
- Common practice: fit $\theta$ to minimize training or validation error
Different covariances: $\Sigma_1 \neq \Sigma_2$

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

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

$$w = \Sigma_1^{-1} \mu_1 - \Sigma_2^{-1} \mu_2$$

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 $\pi_j$, 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 $\text{arg max}_j f_j(x)$.

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

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

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of light, it was the season of darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way – in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

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.