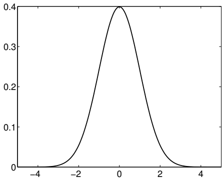
TOOLS: MAXIMUM LIKELIHOOD

Gaussian Distribution

Gaussian density in one dimension

$$g(x;\mu,\sigma) := \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$$

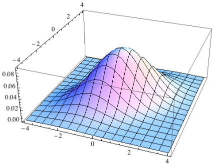


- μ = expected value of x, σ^2 = variance, σ = standard deviation
- The quotient $\frac{x-\mu}{\sigma}$ measures deviation of x from its expected value in units of σ (i.e. σ defines the length scale)

Gaussian density in d dimensions

The quadratic function

$$-\frac{(x-\mu)^2}{2\sigma^2} \quad = \quad -\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)$$



is replaced by a quadratic form:

$$g(\mathbf{x};\boldsymbol{\mu},\Sigma) := \frac{1}{\sqrt{2\pi\det(\Sigma)}}\exp\Big(-\frac{1}{2}\Big\langle(\mathbf{x}-\boldsymbol{\mu}),\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big\rangle\Big)$$

Parametric Models

Models

A **model** $\mathcal{P}$ is a set of probability distributions. We index each distribution by a parameter value $\theta \in \mathcal{T}$; we can then write the model as

$$\mathcal{P} = \{P_\theta | \theta \in \mathcal{T}\} .$$

The set $\mathcal{T}$ is called the **parameter space** of the model.

Parametric model

The model is called **parametric** if the number of parameters (i.e. the dimension of the vector θ) is (1) finite and (2) independent of the number of data points. Intuitively, the complexity of a parametric model does not increase with sample size.

Density representation

For parametric models, we can assume that $\mathcal{T} \subset \mathbb{R}^d$ for some fixed dimension d. We usually represent each P_θ be a density function $p(x|\theta)$.

Maximum Likelihood Estimation

Setting

- Given: Data $x_1, \ldots, x_n$, parametric model $\mathcal{P} = \{p(x|\theta) \mid \theta \in \mathcal{T}\}$.
- Objective: Find the distribution in $\mathcal{P}$ which best explains the data. That means we have to choose a "best" parameter value $\hat{\theta}$.

Maximum Likelihood approach

Maximum Likelihood assumes that the data is best explained by the distribution in $\mathcal{P}$ under which it has the highest probability (or highest density value).

Hence, the **maximum likelihood estimator** is defined as

$$\hat{\theta}_{\text{ML}} := \arg\max_{\theta \in \mathcal{T}} p(x_1, \ldots, x_n|\theta)$$

the parameter which maximizes the joint density of the data.

Analytic Maximum Likelihood

The i.i.d. assumption

The standard assumption of ML methods is that the data is **independent and identically distributed (i.i.d.)**, that is, generated by independently sampling repeatedly from the same distrubtion P.

If the density of P is $p(x|\theta)$, that means the joint density decomposes as

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\theta)$$

Maximum Likelihood equation

The analytic criterion for a maximum likelihood estimator (under the i.i.d. assumption) is:

$$\nabla_\theta \Big(\prod_{i=1}^{n} p(x_i|\theta) \Big) = 0$$

We use the "logarithm trick" to avoid a huge product rule computation.

Logarithm Trick

Recall: Logarithms turn products into sums

$$\log\Big(\prod_i f_i\Big) = \sum_i \log(f_i)$$

Logarithms and maxima

The logarithm is monotonically increasing on $\mathbb{R}_+$.

Consequence: Application of log does not change the *location* of a maximum or minimum:

$$\max_y \log(g(y)) \neq \max_y g(y) \qquad \text{The } \textit{value} \text{ changes.}$$

$$\arg\max_y \log(g(y)) = \arg\max_y g(y) \qquad \text{The } \textit{location} \text{ does not change.}$$

ANALYTIC MLE

Likelihood and logarithm trick

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta) = \arg\max_{\theta} \log\Big(\prod_{i=1}^{n} p(x_i|\theta)\Big) = \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i|\theta)$$

Analytic maximality criterion

$$0 = \sum_{i=1}^{n} \nabla_{\theta} \log p(x_i|\theta) = \sum_{i=1}^{n} \frac{\nabla_{\theta} p(x_i|\theta)}{p(x_i|\theta)}$$

Whether or not we can solve this analytically depends on the choice of the model!

Example: Gaussian Mean MLE

Model: Multivariate Gaussians

The model $\mathcal{P}$ is the set of all Gaussian densities on $\mathbb{R}^d$ with *fixed* covariance matrix Σ,

$$\mathcal{P} = \{g(\,.\,|\mu, \Sigma) \,|\, \mu \in \mathbb{R}^d\} \,,$$

where g is the Gaussian density function. The parameter space is $\mathcal{T} = \mathbb{R}^d$.

MLE equation

We have to solve the maximum equation

$$\sum_{i=1}^{n} \nabla_\mu \log g(x_i|\mu, \Sigma) = 0$$

for μ.

Example: Gaussian Mean MLE

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \nabla_\mu \log \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\Big(-\frac{1}{2}\Big\langle (x_i-\mu), \Sigma^{-1}(x_i-\mu)\Big\rangle\Big) \\
&= \sum_{i=1}^{n} \nabla_\mu \Big(\log\Big(\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\Big) + \log\Big(\exp\Big(-\frac{1}{2}\Big\langle (x_i-\mu), \Sigma^{-1}(x_i-\mu)\Big\rangle\Big)\Big) \\
&= \sum_{i=1}^{n} \nabla_\mu \Big(-\frac{1}{2}\Big\langle (x_i-\mu), \Sigma^{-1}(x_i-\mu)\Big\rangle\Big) = -\sum_{i=1}^{n} \Sigma^{-1}(x_i-\mu)
\end{aligned}
$$

Multiplication by $(-\Sigma)$ gives

$$
0 = \sum_{i=1}^{n} (x_i - \mu) \qquad \Rightarrow \qquad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i
$$

Conclusion

The maximum likelihood estimator of the Gaussian expectation parameter for fixed covariance is

$$
\hat{\mu}_{\text{ML}} := \frac{1}{n}\sum_{i=1}^{n} x_i
$$

Example: Gaussian with Unknown Covariance

Model: Multivariate Gaussians

The model $\mathcal{P}$ is now

$$\mathcal{P} = \{ g(\,.\, | \mu, \Sigma) \,|\, \mu \in \mathbb{R}^d, \Sigma \in \Delta_d \} \;,$$

where Δ_d is the set of positive definite $d \times d$-matrices. The parameter space is $\mathcal{T} = \mathbb{R}^d \times \Delta_d$.

ML approach

Since we have just seen that the ML estimator of μ does not depend on Σ, we can compute $\hat{\mu}_{\text{ML}}$ first. We then estimate Σ using the criterion

$$\sum_{i=1}^{n} \nabla_{\Sigma} \log g(x_i | \hat{\mu}_{\text{ML}}, \Sigma) = 0$$

Solution

The ML estimator of Σ is

$$\hat{\Sigma}_{\text{ML}} := \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{\text{ML}})(x_i - \hat{\mu}_{\text{ML}})^t \;.$$

CLASSIFICATION

ASSUMPTIONS AND TERMINOLOGY

In a **classification problem**, we record measurements $\mathbf{x}_1, \mathbf{x}_2, \dots$.

We assume:

1. All measurements can be represented as elements of a Euclidean $\mathbb{R}^d$.
2. Each $\mathbf{x}_i$ belongs to exactly one out of K categories, called **classes**. We express this using variables $y_i \in [K]$, called **class labels**:

$$y_i = k \quad \Leftrightarrow \quad \text{"}\mathbf{x}_i \text{ in class } k\text{"}$$

3. The classes are characterized by the (unknown!) joint distribution of (X, Y), whose density we denote $p(x, y)$. The conditional distribution with density $p(x|y = k)$ is called the **class-conditional distribution** of class k.
4. The only information available on the distribution p is a set of example measurements *with* labels,

$$(\tilde{\mathbf{x}}_1, \tilde{y}_1), \dots, (\tilde{\mathbf{x}}_n, \tilde{y}_n) \;,$$

called the **training data**.