Introduction to Maximum Likelihood Approaches

November 9, 2020

Table of contents

Acknowledgements

This slide deck borrows heavily from an excellent course on statistical ML by Peter Orbanz.

Other useful resources came from David Blei and Michael Jordan.

Maximum Likelihood

Parametric Models

Models

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

$$\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$$

The set Θ is called the **parameter space** of the model.

Parametric model

The model is called **parametric** if the number of parameters (i.e. 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 $\Theta \subset \mathbb{R}^d$ for some fixed dimension d. We usually represent each P_θ via a density function $p(x \mid \theta)$.

Setting

- Given: Data $x_1, ..., x_n$, parametric model $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$
- Objective: Find the distribution in $\mathcal{P}$ which best explain the data. That means we have to choose a "best" parameter value $\widehat{\theta}$.

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" (technically, the highest density value).

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

$$\widehat{\theta}_{\mathrm{ML}} := \operatorname*{argmax}_{\theta \in \Theta} p(x_1, ..., x_n \mid \theta)$$

the parameter which maximizes the joint density of the data.

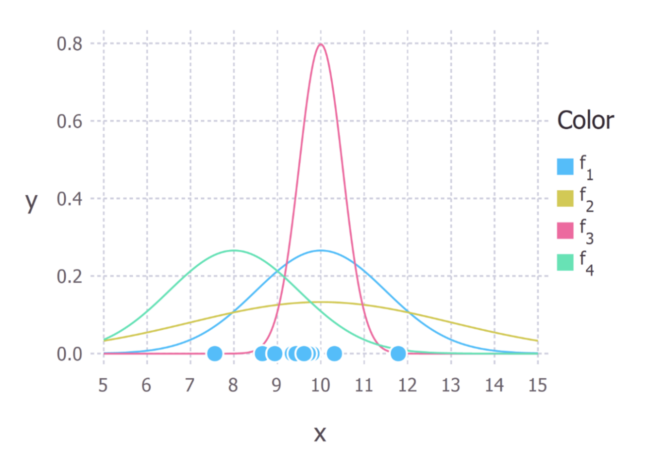
The i.i.d. assumption

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 distribution $\mathcal{P}$.

If the density of $\mathcal{P}$ is $p(x \mid \theta)$, that means the joint density decomposes as

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

Illustration



Ten data points and four possible Gaussians from which they were drawn:
$f_1 \sim \mathcal{N}(10, 2.25)$, $f_2 \sim \mathcal{N}(10, 9)$, $f_3 \sim \mathcal{N}(10, 0.25)$, $f_4 \sim \mathcal{N}(8, 2.25)$.

Image Credit: Jonny Brooks-Bartlett

Maximum Likelihood in practice

Maximum Likelihood equation

In practice, the criterion for a maximum likelihood estimator (under the i.i.d assumption) is

$$\nabla_\theta \left(\prod_{i=1}^{n} p(x_i \mid \theta) \right) = 0$$

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

Logarithm Trick

Recall: Logarithms turn products into sums

$$\log\left(\prod_i f_i\right) = \sum_i \log(f_i)$$

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:

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

$$\operatorname*{argmax}_y \log(g(y)) = \operatorname*{argmax}_y g(y) \qquad \text{The } \textit{location} \text{ does not change.}$$

Likelihood and logarithm trick

$$\widehat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i|\theta) = \underset{\theta}{\operatorname{argmax}} \log \left(\prod_{i=1}^{n} p(x_i|\theta) \right) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(x_i|\theta)$$

Maximum Likelihood in practice (revisited)

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

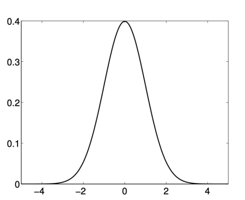
Whether or not we can solve this analytically depends on the choice of model!

Maximum Likelihood Examples

Example: Gaussian Distribution

Gaussian density in one dimension

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



- $\mu =$ expected value of x, $\sigma^2 =$ variance, $\sigma =$ 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).

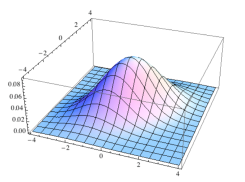
Gaussian density in d dimensions

The quadratic function

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

is replaced by a quadratic form:

$$g(\boldsymbol{x};\boldsymbol{\mu},\Sigma) := \frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$



Example: Gaussian Mean MLE

Model: Multivariate Gausians

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

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

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

MLE equation

We have to solve the maximum likelihood equation

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

for μ.

Example: Gaussian Mean MLE

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \nabla_\mu \log \left[\frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) \right] \\
&= \sum_{i=1}^{n} \nabla_\mu \left[\log \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \right] + \nabla_\mu \left[\log \left(\exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) \right) \right] \\
&= \sum_{i=1}^{n} \nabla_\mu \left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) = -\sum_{i=1}^{n} \Sigma^{-1} (x_i - \boldsymbol{\mu})
\end{aligned}
$$

Multiplication by $(-\Sigma)$ gives

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

Conclusion

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

$$
\widehat{\boldsymbol{\mu}}_{\mathrm{ML}} := \frac{1}{n} \sum_{i=1}^{n} x_i
$$

Example: Gaussian with Unknown Mean, Covariance

Model: Multivariate Gaussians

The model $\mathcal{P}$ is now

$$\mathcal{P} = \{g(\cdot \mid \mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma \in \Delta_d\}$$

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

ML approach

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

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

for μ.

Solution

The ML estimator of Σ is

$$\widehat{\Sigma}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{\mu}_{\text{ML}})(x_i - \widehat{\mu}_{\text{ML}})^T$$

Exercises

Let's split into break-out rooms and try some ML exercises

Anomaly Detection with Multivariate Gaussians

Anomaly Detection with Multivariate Gaussians

Given a fitted Gaussian model, how can we assess the anomalousness of test data?

Anomaly Detection with Multivariate Gaussians

Given a fitted Gaussian model, how can we assess the anomalousness of test data?

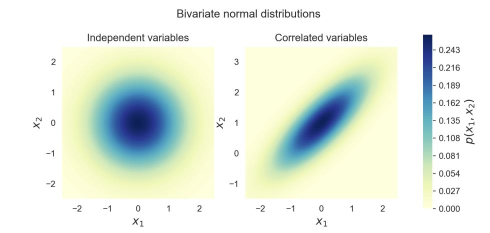


Image Credit: Peter Roelants

Anomaly Detection with Multivariate Gaussians

Given a fitted Gaussian model, how can we assess the anomalousness of test data?

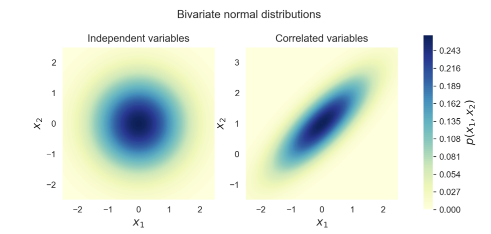


Image Credit: Peter Roelants

Mahalanobis Distance

For a Gaussian random variable $X \sim N(\boldsymbol{\mu}, \Sigma)$, the quadratic form (or *squared Mahalanobis distance*) has known distribution

$$\Delta^2 = (\boldsymbol{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2(d)$$

This can be used to assess the anomalousness of test data.

Exercises

Let's split into break-out rooms and try some exercises

Optimization

Optimization problem

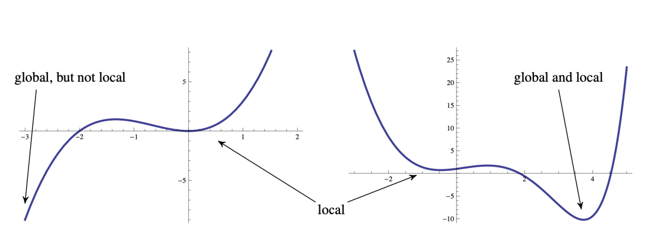
An *optimization problem* for a given function $f : \mathbb{R}^d \to \mathbb{R}$ is a problem of the form

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

which we read as "find $\boldsymbol{x}_0 = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$"'.

Note that finding the maximum likelihood requires minimizing the cost function that is the negative log likelihood.

Local and global modes



Local and global minima

A minimum of a function f at x is called

- **Global** if f assumes no smaller value on its domain.
- **Local** if there is some open neighborhood U of x such that $f(x)$ is a global minimum of f restricted to U.

Image Credit: Peter Orbanz

Analytic criteria for local minima

Recall that $\boldsymbol{x}$ is a local minimum of f if

$$f'(\boldsymbol{x}) = 0 \quad \text{and} \quad f''(\boldsymbol{x}) > 0$$

In $\mathbb{R}^d$,

$$\nabla f(\boldsymbol{x}) = 0 \text{ and } H_f(\boldsymbol{x}) = \left(\frac{\delta f}{\delta x_i \delta x_j}(\boldsymbol{x})\right)_{i,j=1,\ldots,n} \text{ positive definite}$$

The $d \times d$-matrix $H_f(\boldsymbol{x})$ is called the **Hessian matrix** of f at $\boldsymbol{x}$.

The MLE and Global Maximizers

You may have noticed that the maximum likelihood equation is only tracking a *local* maximality criterion. In fact, it also ignored the second-order condition. What gives?

- Many well-known distributions[1] have strictly concave likelihoods, in which case the MLE equation is sufficient to verify a global maximum.
- For many other distributions, it can be hard to find the global maximizer of the likelihood. Thus a local maximizer is often used and is called an MLE. The local optimizer is typically found by an optimization procedure, from which the second order condition generally follows.

---

[1] In particular, those in the exponential family. We will cover this in the next slide deck.

Demo

Gradient Ascent Demo

Exponential Family

Definition

We consider a model $\mathcal{P}$ for data in a sample space $\mathcal{X}$ with parameter space $\Theta \subset \mathbb{R}^m$. Each distribution in $\mathcal{P}$ has density $p(x \mid \theta)$ for some $\theta \in \Theta$.

The model is called an **exponential family model** (EFM) if p can be written as

$$p(x \mid \theta) = h(x) \exp\{\eta(\theta)^T s(x) - a(\theta)\}$$

where we refer to

- h as the base measure
- η as the natural parameter
- s as the sufficient statistics
- a as the log normalizer.

Exponential families are important because

- Many important parametric models (Gaussian, Poisson, beta, gamma, etc.) are EFM's.
- The special form of p gives them many nice properties. For example, exponential family likelihoods are *strictly convex.*[2]
- Many algorithms and methods can be formulated generically for all EFM's.

An observation

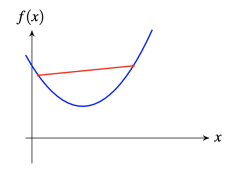
The data and the parameter interact only through the linear term $\eta(\theta)^T s(x)$ in the exponent.

[2] And for why we care about *that*, see the next slide. Other useful properties will come up as we go along.

Convex Functions

Definition

A function f is **convex** if every line segment between function values lies above the graph of f



Analytic criterion

A twice differentiable function is convex if $f''(\boldsymbol{x}) \geq 0$ (or $H_f(\boldsymbol{x})$ positive semidefinite) for all $\boldsymbol{x}$.

Implications for optimization

If f is convex, then:

- $f'(\boldsymbol{x}) = 0$ is a sufficient criterion for a minimum.
- Local minima are global.
- If f is strictly convex ($f'' > 0$ or H_f positive definite), there is only one minimum (which is both global and local).

Example: The Gaussian distribution

The familiar form of the univariate Gaussian is

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

Example: The Gaussian distribution

The familiar form of the univariate Gaussian is

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

We put it in exponential family form by expanding the square

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 - \log\sigma\right)$$

which reveals the exponential family where

$$\begin{aligned} \eta &= [\mu/\sigma^2, -1/2\sigma^2] \\ s(x) &= [x, x^2] \\ a(\eta) &= \mu^2/2\sigma^2 + \log\sigma \\ h(x) &= 1/\sqrt{2\pi} \end{aligned}$$

Example: The Bernoulli distribution

As an example, let's put the Bernoulli (in its usual form) into exponential family form. The Bernoulli you are used to seeing is:

$$p(x \mid \pi) = \pi^x \, (1 - \pi)^{1-x} \quad x \in \{0, 1\}$$

Example: The Bernoulli distribution

As an example, let's put the Bernoulli (in its usual form) into exponential family form. The Bernoulli you are used to seeing is:

$$p(x \mid \pi) = \pi^x \, (1-\pi)^{1-x} \quad x \in \{0,1\}$$

In exponential family form:

$$\begin{aligned}
p(x \mid \pi) &= \exp\left(\log\left[\pi^x \, (1-\pi)^{1-x}\right]\right) \\
&= \exp\Big(x \log \pi + (1-x)\log(1-\pi)\Big) \\
&= \exp\Big(x \log \pi - x\log(1-\pi) + \log(1-\pi)\Big) \\
&= \exp\Big(x \log(\pi/(1-\pi)) + \log(1-\pi)\Big)
\end{aligned}$$

which reveals the exponential family where

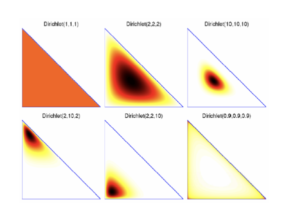
$$\begin{aligned}
\eta &= \log(\pi/(1-\pi)) \\
s(x) &= x \\
a(\eta) &= -\log(1-\pi) = \log(1+e^\eta) \\
h(x) &= 1
\end{aligned}$$

Note that the relationship between π and η is invertible

$$\pi = 1/(1+e^{-\eta})$$

This it the *logistic function*.

Example: The Dirichlet Distribution



We can write the density of the Dirichlet distribution in exponential form:

$$
\begin{aligned}
p(\pi \mid \alpha) &= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} \\
&= \exp\left\{ \sum_{k=1}^{K} (\alpha_k - 1) \log \pi_k - \left[\sum_k \log \Gamma(\alpha_k) - \log \Gamma(\sum \alpha_k) \right] \right\}
\end{aligned}
$$

with natural parameter $\eta(\alpha) = [\alpha_1 - 1, ..., \alpha_K - 1]^T$, sufficient statistics $s(\pi) = \log \pi = [\log \pi_1, ..., \log \pi_K]^T$, base measure $h(\pi) = 1$, and log normalizer $a(\alpha) = \sum_k \log \Gamma(\alpha_k) - \log \Gamma(\sum_k \alpha_k)$. $\square$

Examples of Exponential Families

Model	Sample space	Sufficient statistic
Gaussian	$\mathbb{R}^d$	$S(\mathbf{x}) = (\mathbf{x}\mathbf{x}^t, \mathbf{x})$
Gamma	$\mathbb{R}_+$	$S(x) = (\ln(x), x)$
Poisson	$\mathbb{N}_0$	$S(x) = x$
Multinomial	$\{1, \dots, K\}$	$S(x) = x$
Wishart	Positive definite matrices	(requires more details)
Mallows	Rankings (permutations)	(requires more details)
Beta	$[0, 1]$	$S(x) = (\ln(x), \ln(1 - x))$
Dirichlet	Probability distributions on d events	$S(\mathbf{x}) = (\ln x_1, \dots, \ln x_d)$
Bernoulli	$\{0, 1\}$	$S(x) = x$
...	...	...

Roughly speaking

On every sample space, there is a "natural" statistic of interest. On a space with Euclidean distance, for example, it is natural to measure both location and correlation; on categories (which have no "distance" from each other), it is more natural to measure only expected numbers of counts.

The Exponential Family and Maximum Likelihood

i.i.d samples from an exponential family distribution

If $\boldsymbol{x} = (x_1, ..., x_n)$ are n independent samples from the same exponential family distribution, then

$$p(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} h(x_i) \exp\left\{\eta(\theta)^T \sum_{i=1}^{n} s(x_i) - n\, a(\eta(\theta))\right\}$$

Maximum likelihood with exponential families

The goal for maximum likelihood is to determine parameter

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \log p(\boldsymbol{x} \mid \theta)$$

Let us assume that $\boldsymbol{x} = (x_1, ..., x_n)$ are i.i.d observations from a fixed exponential family, so that the likelihood has form above.

The Exponential Family and Maximum Likelihood

Let us compute the gradient with respect to the natural parameter η of $\ell(\eta) := \log p(\boldsymbol{x} \mid \eta)$

$$\nabla_\eta \ell(\eta) = \sum_{i=1}^{n} s(x_i) - n \, \nabla_\eta a(\eta)$$

Setting the gradient to zero, we obtain

$$\nabla_\eta a(\eta) = \frac{1}{n} \sum_{i=1}^{n} s(x_i)$$

But[3] $\nabla_\eta a(\eta) = \mathbb{E}[s(X)]$. Thus, we should set θ_{ML} such that

$$\mu(\theta_{ML}) = \frac{1}{n} \sum_{i=1}^{n} s(x_i)$$

where $\mu := \mathbb{E}[s(x)]$ refers to the mean parametrization of the likelihood.

[3] A useful fact about exponential families. The proof is straightforward.