

Maximum likelihood estimation



École des Ponts
ParisTech

Guillaume Obozinski

Ecole des Ponts - ParisTech



Master MVA

Outline

- 1 Statistical concepts
- 2 A short review of convex analysis and optimization
- 3 The maximum likelihood principle

Statistical concepts

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \{p_\theta(x) \mid \theta \in \Theta\}$$

Bernoulli model: $X \sim \text{Ber}(\theta)$ $\Theta = [0, 1]$

$$p_\theta(x) = \theta^x (1 - \theta)^{(1-x)}$$

Binomial model: $X \sim \text{Bin}(n, \theta)$ $\Theta = [0, 1]$

$$p_\theta(x) = \binom{n}{x} \theta^x (1 - \theta)^{(n-x)}$$

Multinomial model: $X \sim \mathcal{M}(n, \pi_1, \pi_2, \dots, \pi_K)$ $\Theta = [0, 1]^K$

$$p_\theta(x) = \binom{n}{x_1, \dots, x_k} \pi_1^{x_1} \dots \pi_k^{x_k}$$

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \dots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \dots, Y_K)^\top$ with

$$Y_k = 1_{\{C=k\}}$$

For example if $K = 5$ and $c = 4$ then $\mathbf{y} = (0, 0, 0, 1, 0)^\top$.
So $\mathbf{y} \in \{0, 1\}^K$ with $\sum_{k=1}^K y_k = 1$.

$$\mathbb{P}(C = k) = \mathbb{P}(Y_k = 1) \quad \text{and} \quad \mathbb{P}(Y = \mathbf{y}) = \prod_{k=1}^K \pi_k^{y_k}.$$

Bernoulli, Binomial, Multinomial

$Y \sim \text{Ber}(\pi)$	$(Y_1, \dots, Y_K) \sim \mathcal{M}(1, \pi_1, \dots, \pi_K)$
$p(y) = \pi^y (1 - \pi)^{1-y}$	$p(\mathbf{y}) = \pi_1^{y_1} \dots \pi_K^{y_K}$
$N_1 \sim \text{Bin}(n, \pi)$	$(N_1, \dots, N_K) \sim \mathcal{M}(n, \pi_1, \dots, \pi_K)$
$p(n_1) = \binom{n}{n_1} \pi^{n_1} (1 - \pi)^{n-n_1}$	$p(\mathbf{n}) = \binom{n}{n_1 \dots n_K} \pi_1^{n_1} \dots \pi_K^{n_K}$

with

$$\binom{n}{i} = \frac{n!}{(n-i)!i!} \quad \text{and} \quad \binom{n}{n_1 \dots n_K} = \frac{n!}{n_1! \dots n_K!}$$

Gaussian model

Scalar Gaussian model : $X \sim \mathcal{N}(\mu, \sigma^2)$

X real valued r.v., and $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

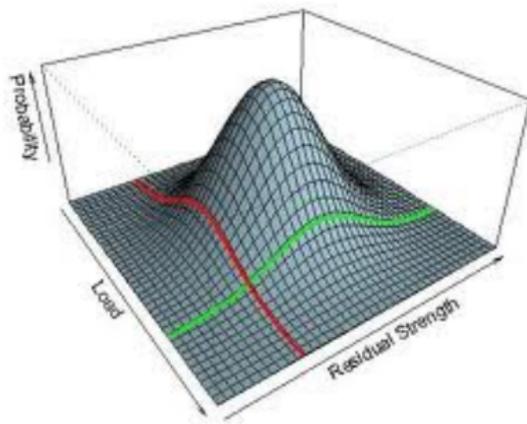
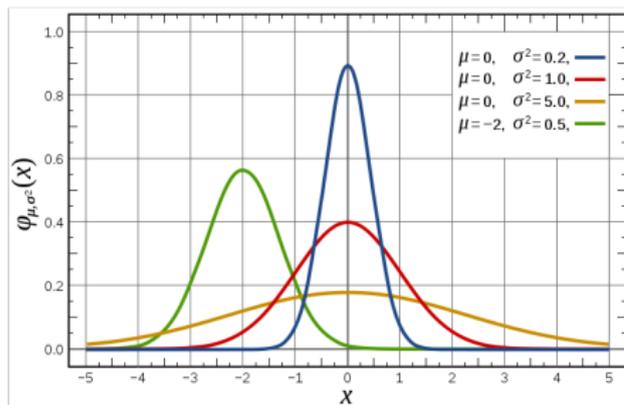
$$p_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right)$$

Multivariate Gaussian model: $X \sim \mathcal{N}(\mu, \Sigma)$

X r.v. taking values in \mathbb{R}^d . If \mathcal{K}_d is the set of positive definite matrices of size $d \times d$, and $\theta = (\mu, \Sigma) \in \Theta = \mathbb{R}^d \times \mathcal{K}_d$.

$$p_{\mu, \Sigma}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right)$$

Gaussian densities



A short review of convex analysis and optimization

Review: convex analysis

Convex function

$$\forall \lambda \in [0, 1], \quad f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$

Strictly convex function

$$\forall \lambda \in]0, 1[, \quad f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) < \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$

Strongly convex function

$$\exists \mu > 0, \text{ s.t. } \mathbf{x} \mapsto f(\mathbf{x}) - \mu \|\mathbf{x}\|^2 \text{ is convex}$$

Equivalently:

$$\forall \lambda \in [0, 1], \quad f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) - \mu \lambda (1 - \lambda) \|\mathbf{x} - \mathbf{y}\|^2$$

The largest possible μ is called the strong convexity constant.

Minima of convex functions

Proposition (Supporting hyperplane)

If f is convex and differentiable at \mathbf{x} then

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

Convex function

All local minima are global minima.

Strictly convex function

If there is a local minimum, then it is unique and global.

Strongly convex function

There exists a unique local minimum which is also global.

Minima and stationary points of differentiable functions

Definition (Stationary point)

For f differentiable, we say that \mathbf{x} is a stationary point if $\nabla f(\mathbf{x}) = 0$.

Theorem (Fermat)

If f is differentiable at \mathbf{x} and \mathbf{x} is a local minimum, then \mathbf{x} is stationary.

Theorem (Stationary point of a convex differentiable function)

If f is convex and differentiable at \mathbf{x} and \mathbf{x} is stationary then \mathbf{x} is a minimum.

Theorem (Stationary points of a twice differentiable functions)

For f twice differentiable at \mathbf{x}

- if \mathbf{x} is a local minimum then $\nabla f(\mathbf{x}) = 0$ and $\nabla^2 f(\mathbf{x}) \succeq 0$.*
- conversely if $\nabla f(\mathbf{x}) = 0$ and $\nabla^2 f(\mathbf{x}) \succ 0$ then \mathbf{x} is a strict local minimum.*

Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \dots, X^{(n)}$$

A common assumption is that the variables are **i.i.d.**

- **independent**
- **identically distributed**, i.e. have the same distribution P .

This collection of observations is called

- the *sample* or the *observations* in statistics
- the *samples* in engineering
- the *training set* in machine learning

The maximum likelihood principle

Maximum likelihood principle

- Let $\mathcal{P}_\Theta = \{p(x; \theta) \mid \theta \in \Theta\}$ be a *model*
- Let x be an observation

Likelihood:

$$\begin{aligned}\mathcal{L} : \Theta &\rightarrow \mathbb{R}_+ \\ \theta &\mapsto p(x; \theta)\end{aligned}$$

Maximum likelihood estimator:

$$\hat{\theta}_{\text{ML}} = \operatorname{argmax}_{\theta \in \Theta} p(x; \theta)$$

Case of i.i.d data

If $(x_i)_{1 \leq i \leq n}$ is an i.i.d. sample of size n :

$$\hat{\theta}_{\text{ML}} = \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n p_\theta(x_i) = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \log p_\theta(x_i)$$



Sir Ronald Fisher
(1890-1962)

The maximum likelihood estimator

The MLE

- does not always exist
- is not necessarily unique
- is not *admissible* in general

MLE for the Bernoulli model

Let X_1, X_2, \dots, X_n an i.i.d. sample $\sim \text{Ber}(\theta)$. The log-likelihood is

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^n \log p(x_i; \theta) = \sum_{i=1}^n \log [\theta^{x_i} (1 - \theta)^{1-x_i}] \\ &= \sum_{i=1}^n (x_i \log \theta + (1 - x_i) \log(1 - \theta)) = N \log(\theta) + (n - N) \log(1 - \theta)\end{aligned}$$

with $N := \sum_{i=1}^n x_i$.

- $\theta \mapsto \ell(\theta)$ is strongly concave \Rightarrow the MLE exists and is unique.
- since ℓ differentiable + strongly concave its maximizer is the unique stationary point

$$\nabla \ell(\theta) = \frac{\partial}{\partial \theta} \ell(\theta) = \frac{N}{\theta} - \frac{n - N}{1 - \theta}.$$

Thus

$$\hat{\theta}_{\text{MLE}} = \frac{N}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

MLE for the multinomial

Done on the board. See lecture notes.

Brief review of Lagrange duality

Convex optimization problem with linear constraints

For

- f a convex function,
- $\mathcal{X} \subset \mathbb{R}^p$ a convex set included in the domain of f ,
- $\mathbf{A} \in \mathbb{R}^{n \times p}$, $\mathbf{b} \in \mathbb{R}^n$,

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{Ax} = \mathbf{b} \quad (P)$$

Lagrangian

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T (\mathbf{Ax} - \mathbf{b})$$

with $\boldsymbol{\lambda} \in \mathbb{R}^n$ the *Lagrange multiplier*.

Properties of the Lagrangian

Link between primal and Lagrangian

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^n} L(\mathbf{x}, \boldsymbol{\lambda}) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{Ax} = \mathbf{b} \\ +\infty & \text{otherwise.} \end{cases}$$

So that

$$\min_{\mathbf{x} \in \mathcal{X}: \mathbf{Ax} = \mathbf{b}} f(\mathbf{x}) = \min_{\mathbf{x} \in \mathcal{X}} \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} L(\mathbf{x}, \boldsymbol{\lambda})$$

Lagrangian dual objective function

$$g(\boldsymbol{\lambda}) = \min_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda})$$

Dual optimization problem

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^n} g(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \min_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}) \quad (D)$$

Maxmin-minmax inequality, weak and strong duality

For any $f : \mathbb{R}^n \times \mathbb{R}^m$ and any $w \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$, we have

$$\max_{z \in Z} \min_{w \in W} f(w, z) \leq \min_{w \in W} \max_{z \in Z} f(w, z).$$

Weak duality

$$d^* := \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} g(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} \min_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}) \leq \min_{\mathbf{x} \in \mathcal{X}} \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} L(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) =: p^*$$

So that in general, we have $d^* \leq p^*$. This is called weak duality

Strong duality

In some cases, we have strong duality:

- $d^* = p^*$
- Solutions to (P) and (D) are the same

Slater's qualification condition

Slater's qualification condition is a condition on the constraints of a convex optimization problem that guarantees that strong duality holds.

For linear constraints, Slater's condition is very simple:

Slater's condition for a cvx opt. pb with lin. constraints

If there exists an \mathbf{x} in the relative interior of $\mathcal{X} \cap \{\mathbf{Ax} = \mathbf{b}\}$ then strong duality holds.

MLE for the univariate and multivariate Gaussian

Done on the board. See lecture notes.