Machine Learning

Section 7: More on distributions, models, MAP, ML

关于分布、模型、MAP、ML的更多信息

Stefan Harmeling

13. October 2021

Gaussian distribution

Univariate Gaussian distribution

see MLPP 2.4.1 (Murphy: Machine Learning: a Probabilistic Perspective)

- random variable X is real-valued
- parameters μ called mean, $\sigma^2 > 0$ called variance
- X has univariate Gaussian distribution, written

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

- probability density function

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Multivariate Gaussian distribution

see MLPP 2.5.2

- random vector X has real-valued components
- parameters μ called mean vector, pos-def symmetric matrix Σ called covariance
- X has multivariate Gaussian distribution, written

$$X \sim \mathcal{N}(\mu, \Sigma)$$

- probability density function

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

- special case: $\mathcal{N}(\mu, \sigma^2)$

Closed under sum- and product rule:

A Gaussian joint distribution

$$p(x, y) = \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} \mu \\ \nu \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right)$$

has Gaussian marginals

$$p(x) = \int p(x, y)\, dy = \mathcal{N}(x, \mu, A)$$
$$p(y) = \int p(x, y)\, dx = \mathcal{N}(y, \nu, C)$$

and Gaussian conditionals

$$p(x|y) = p(x, y)/p(y) = \mathcal{N}(x, \mu + BC^{-1}(y - \nu), A - BC^{-1}B^T)$$
$$p(y|x) = p(x, y)/p(x) = \mathcal{N}(y, \nu + B^T A^{-1}(x - \mu), C - B^T A^{-1}B)$$

Important non-Gaussian distributions
重要的非高斯分布

Binomial distribution

see MLPP 2.3.1

- toss a coin n times
- let random variable $X \in \{0, \dots, n\}$ be number of heads
- let θ be the probabilty of heads
- X has binomial distribution, written

$$X \sim \mathrm{Bin}(n, \theta)$$

- probability mass function

$$\mathrm{Bin}(k|n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

- mean $= n\theta$, var $= n\theta(1 - \theta)$

Bernoulli distribution

see MLPP 2.3.1

- toss a coin once
- let random variable $X \in \{0, 1\}$ be a binary variable
- let θ be the probabilty of heads
- X has Bernoulli distribution, written

$$X \sim \mathrm{Ber}(\theta)$$

- probability mass function

$$\mathrm{Ber}(x|\theta) = \theta^{[x=1]}(1-\theta)^{[x=0]} = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

  using Iverson brackets $[A] = 1$ if A is true, zero otherwise
- mean $= \theta$, var $= \theta(1-\theta)$
- special case: $\mathrm{Ber}(\theta) = \mathrm{Bin}(1, \theta)$

Multinomial distribution

see MLPP 2.3.2

- toss a K-sided dice n times
- let $X = (x_1, \ldots, x_K)$ be a random vector, with x_j being the number of times side j occurs, $\sum_j x_j = n$
- let $\theta = (\theta_1, \ldots, \theta_K)$ be the parameter vector, with $\sum_j \theta_j = 1$ and $\theta_j \geq 0$
- θ_j be the probabilty of side j of the dice
- X has multinomial distribution, written

$$X \sim \mathrm{Mu}(n, \theta)$$

- probability mass function

$$\mathrm{Mu}(x|n, \theta) = \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

with multinomial coefficient

$$\binom{n}{x_1 \ldots x_K} = \frac{n!}{x_1! x_2! \cdots x_K!}$$

Multinoulli distribution

see MLPP 2.3.2

- toss a K-sided dice once
- let $X = (x_1, \ldots, x_K)$ be a random vector, with x_j being binary, such that only one is non-zero
- let $\theta = (\theta_1, \ldots, \theta_K)$ be the parameter vector, with $\sum_j \theta_j = 1$ and $\theta_j \geq 0$
- θ_j be the probabilty of side j of the dice
- X has multinoulli distribution, written

$$X \sim \mathrm{Cat}(\theta) = \mathrm{Mu}(1, \theta)$$

- probability mass function

$$\mathrm{Cat}(x|\theta) = \prod_{j=1}^{K} \theta_j^{x_j}$$

- aka categorical or discrete distribution

Tossing dice (1)

- tossing n times a K sided dice
- let X be random vector of number of times side j appeared
- distribution of X: Multinomial

$$X \sim \mathrm{Mu}(n, \theta)$$

  with parameter vector θ
- assume $n = 1$: Multinoulli

$$\mathrm{Cat}(\theta) = \mathrm{Mu}(1, \theta)$$

- assume case $K = 2$: Binomial

$$\mathrm{Bin}(n, \theta) = \mathrm{Mu}(n, (\theta, 1 - \theta))$$

  with $\theta \in [0, 1]$
- assume $n = 1$ and $K = 2$: Bernoulli

$$\mathrm{Ber}(\theta) = \mathrm{Bin}(1, \theta) = \mathrm{Mu}(1, (\theta, 1 - \theta)) = \mathrm{Cat}((\theta, 1 - \theta))$$

  with $\theta \in [0, 1]$

Tossing dice (2)

	$n = 1$	n>1
$k = 2$	Bernoulli	Binomial
$k > 2$	Multinoulli	Multinomial

Poisson distribution

see MLPP 2.3.3

- counts of rare events
- let random variable $X \in \{0, 1, \dots\}$ be the number of events in some time interval
- let $\lambda > 0$ be the parameter (the rate)
- X has Poisson distribution, written

$$X \sim \mathrm{Poi}(\lambda)$$

- probability mass function

$$\mathrm{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

- e.g. number of emails you receive every days is Poisson distributed

Beta distribution

see MLPP 2.4.6

- random variable $\theta \in [0, 1]$ (interval between zero and one)
- parameters $a > 0$ and $b > 0$
- θ has beta distribution, written

$$\theta \sim \mathrm{Beta}(a, b)$$

- probability density function

$$\mathrm{Beta}(\theta | a, b) = \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1}$$

with $B(a, b)$ being the beta function

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

- mean $= a/(a + b)$, mode $= (a - 1)/(a + b - 2)$

Gamma function, Beta function, and all that

from http://en.wikipedia.org/wiki/Gamma_function
and http://en.wikipedia.org/wiki/Beta_function

Gamma function (extension of factorial function)

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, dt \qquad \text{for } z \in \mathbb{C}$$

$$\Gamma(n) = (n-1)! = n!/n \qquad \text{for } n \in \mathbb{N}$$

Beta function (extension of . . . ?)

$$\begin{aligned}
B(x,y) &= \int_0^1 t^{x-1}(1-t)^{y-1}\, dt \\
&= \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} && \text{for } x,y \in \mathbb{C} \text{ with } x+\bar{x}, y+\bar{y} > 0 \\
B(m,n) &= \frac{(m-1)!\,(n-1)!}{(m+n-1)!} && \text{for } m,n \in \mathbb{N} \\
&= \binom{m+n}{n}^{-1} \frac{m+n}{m\,n} && \text{binomial coefficient}
\end{aligned}$$

Dirichlet distribution

see MLPP 2.5.4

- random vector $\theta = (\theta_1, \ldots, \theta_K)$ with values in probability simplex, i.e. $\sum_j \theta_j = 1$, $\theta_j \geq 0$.
- parameter vector $\alpha = (\alpha_1, \ldots, \alpha_K)$, with $\alpha_j > 0$
- θ has Dirichlet distribution, written

$$\theta \sim \mathrm{Dir}(\alpha)$$

- probability density function

$$\mathrm{Dir}(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

  with $B(\alpha)$ generalizing the beta function

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

- special case: $\mathrm{Beta}(a, b) = \mathrm{Dir}((a, b))$

Again tossing coins and dice

Throw a coin ($k = 2$) or a dice $k > 2$).

Distributions for the outcome

- coin ($k = 2$): $X \sim \mathrm{Ber}(\theta)$ with θ being scalar
- dice ($k > 2$): $X \sim \mathrm{Mu}(\theta)$ with θ being vector (length k)

Distributions for the parameter (conjugate priors!)

- coin ($k = 2$): $\theta \sim \mathrm{Beta}(a, b)$ with a and b being scalar
- dice ($k > 2$): $\theta \sim \mathrm{Dir}(\alpha)$ with α being vector (length k)

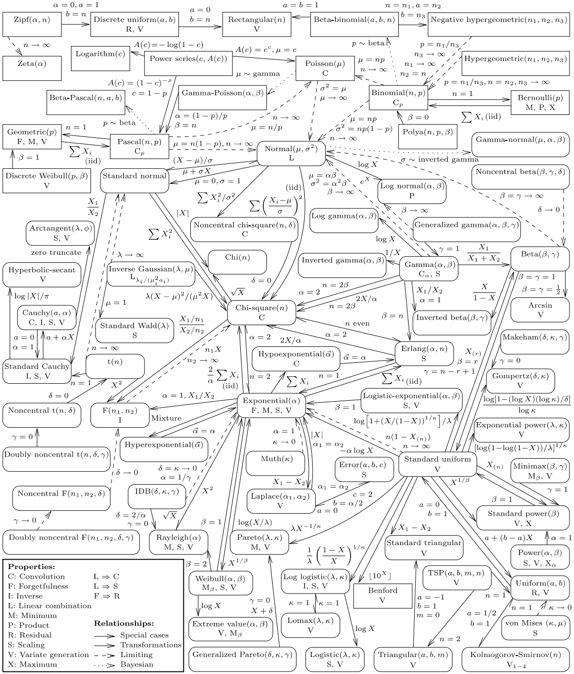

Figure 1. Univariate distribution relationships.

previous graphics from: “Univariate Distribution Relationships”, Lawrence M. Leemis and Jacquelyn T. McQueston, The American Statistician, February 2008, Vol. 62, No. 1, page 47

Beta-binomial model

MLPP 3.3

Data

- flip repeatedly a coin with unknown heads probability θ
- k number of heads, n total number of throws
- k is the data $\mathcal{D}$
- same as wearing glasses example (lecture 03)

Specify

$$\begin{aligned} \theta &\sim \mathrm{Beta}(a, b) & p(\theta) &= \mathrm{Beta}(\theta|a, b) & &\text{prior} \\ k|\theta &\sim \mathrm{Bin}(n, \theta) & p(k|\theta) &= \mathrm{Bin}(k|n, \theta) & &\text{likelihood} \end{aligned}$$

Infer

$$\begin{aligned} \theta|k &\sim \mathrm{Beta}(a+k, b+n-k) & &\text{posterior} \\ p(\theta|k) &= \mathrm{Beta}(\theta|a+k, b+n-k) & &\text{posterior} \end{aligned}$$

- both notations are fine: $\theta \sim \mathrm{Beta}(a, b)$ and $p(\theta) = \mathrm{Beta}(\theta|a, b)$

What can we do with the posterior?

How can I get a point estimate?
我如何能得到一个点的估计？

Summarize the posterior: MAP vs ML

- let's denote the data as $\mathcal{D}$ (was k on the previous slide)
- summarize the posterior by a point estimate
- maximum a posteriori estimate (MAP)

$$\theta_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta)$$

  (aka mode of the posterior)
- similar to maximum likelihood (ML) estimate

$$\theta_{\text{ML}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

- likelihood term dominates for lots of data, thus the data overwhelms the prior and MAP converges against ML
- MAP and ML ignore variance of posterior
- nonetheless, MAP is useful if the posterior is peaked, ML useful if we have lots of data

Famous ML estimator for Gaussian likelihoods

Setup

- consider Gaussian distributed data points $X_1, \ldots, X_n \sim \mathcal{N}(x|\mu, I)$
- goal: estimate mean μ

Maximize the likelihood (aka ML)

$$
\begin{aligned}
\mu_{\mathrm{ML}} &= \arg\max_{\mu} p(X_1, \ldots, X_n|\mu) \\
&= \arg\max_{\mu} \log p(X_1, \ldots, X_n|\mu) \\
&= \arg\max_{\mu} \log \prod_{i=1}^{n} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x_i-\mu)^T(x_i-\mu)} \\
&= \arg\max_{\mu} \sum_{i=1}^{n} \log e^{-\frac{1}{2}(x_i-\mu)^T(x_i-\mu)} \\
&= \arg\min_{\mu} \sum_{i=1}^{n} \|x_i - \mu\|^2
\end{aligned}
$$

Thus we derived the method of *least-squares*!

Posterior predictive distribution

Alternative to point estimates such as ML and MAP:

- posterior expresses our belief state about the world

$$p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|a+k, b+n-k)$$

- use it to make predictions! (scientific method)
- define posterior predictive distribution

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1,\theta|\mathcal{D})\,d\theta = \int_0^1 p(x=1|\theta)\,p(\theta|\mathcal{D})\,d\theta$$

  where x is e.g. a random variable for the outcome of a future coin toss, note that $x \perp\!\!\!\perp \mathcal{D} \mid \theta$
- posterior predictive distribution integrates out the unknown parameter using the posterior

Back to the beta-binomial model

- MAP and ML

$$
\begin{aligned}
\theta_{\text{MAP}} &= \arg\max_{\theta} p(\theta|\mathcal{D}) \\
&= \arg\max_{\theta} \text{Beta}(\theta, a+k, b+n-k) = \frac{a+k-1}{a+b+n-2} \\
\theta_{\text{ML}} &= \arg\max_{\theta} p(\mathcal{D}|\theta) = \arg\max_{\theta} \text{Bin}(k|n,\theta) = \frac{k}{n}
\end{aligned}
$$

- ML equals the MAP estimate for uniform prior on θ, i.e. for $a = 1$, $b = 1$.
- posterior predictive distribution

$$
\begin{aligned}
p(x=1|\mathcal{D}) &= \int_0^1 p(x=1|\theta) p(\theta|\mathcal{D}) d\theta \\
&= \int_0^1 \theta \, \text{Beta}(\theta|a+k, b+n-k) d\theta \\
&= \frac{a+k}{a+b+n} = \text{ posterior mean}
\end{aligned}
$$

Which should I choose? (1)

MLPP 5.7

Bayesian decision theory

- turn priors into posteriors to update your beliefs
- how to convert beliefs into actions?
- define a *loss function* which tells us how expensive it is to be wrong
- i.e. what is the loss $L(\hat{\theta}, \theta)$ if we pick parameter $\hat{\theta}$ while θ is the true one
- given the posterior $p(\theta|\mathcal{D})$ pick the $\hat{\theta}$ that minimizes the *posterior expected loss*

$$\rho(\hat{\theta}) = \int L(\hat{\theta}, \theta) p(\theta|\mathcal{D}) d\theta$$

- *Bayes estimator*, aka *Bayes decision rule*

$$\hat{\theta} = \arg\min_{\hat{\theta}} \rho(\hat{\theta})$$

Which should I choose? (2)

MLPP 5.7

Some common loss functions

- for the 0-1 loss

$$L(\hat{\theta}, \theta) = \begin{cases} 0 & \text{if } \hat{\theta} = \theta \\ 1 & \text{if } \hat{\theta} \neq \theta \end{cases}$$

  the Bayes estimator is MAP
- for the quadratic loss, aka l_2 loss, aka squared error

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

  the Bayes estimator is the posterior mean
- for the robust loss, aka absolute error, aka l_1 loss

$$L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$$

  the Bayes estimator is the posterior median

Which should I choose? (3)

Story:

> *You are at the NeurIPS confernce in a big hotel, standing in front of five elevators. Where should stand to minimize the length of the way to the next open elevator?*

What loss function should you use? What is the resulting estimator?

What else can we do with the posteriors?

Don’t we usually just want point estimates?

Inference for a difference in proportions

MLPP 5.2.3, see link in MLPP for the source

Story

Two sellers at Amazon have the same price. One has 90 positive, 10 negative reviews. The other one 2 positive, 0 negative. Who should you buy from?

Apply two beta-binomial models (assuming uniform priors)

$$p(\theta_1|\mathcal{D}_1) = \mathrm{Beta}(\theta_1|91, 11) \qquad \text{posterior about reliability}$$
$$p(\theta_2|\mathcal{D}_2) = \mathrm{Beta}(\theta_2|3, 1) \qquad \text{posterior about reliability}$$

Compute probability that seller 1 is more reliable than seller 2:

$$p(\theta_1 > \theta_2|\mathcal{D}_1, \mathcal{D}_2)$$
$$= \int_0^1 \int_0^1 [\theta_1 > \theta_2]\, \mathrm{Beta}(\theta_1|91, 11)\, \mathrm{Beta}(\theta_2|3, 1) d\theta_1 d\theta_2 \approx 0.710$$

using numerical integration (your exercise...).

Beta-binomial model

MLPP 3.3

Data

- flip repeatedly a coin with unknown heads probability θ
- k number of heads, n total number of throws
- k is the data $\mathcal{D}$
- same as wearing glasses example (lecture 03)

Specify

$$
\begin{aligned}
p(\theta) &= \mathrm{Beta}(\theta|a, b) && \text{prior} \\
p(\mathcal{D}|\theta) &= \mathrm{Bin}(k|n, \theta) && \text{likelihood}
\end{aligned}
$$

Infer

$$
p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|a + k, b + n - k) \qquad \text{posterior}
$$

Dirichlet-multinomial model

MLPP 3.4

Data

- throw n times a dice with unknown probabilities $\theta = (\theta_1, \ldots, \theta_K)$
- data $\mathcal{D} = (x_1, \ldots, x_K)$, with x_j number of times side j

Specify

$$
\begin{aligned}
p(\theta) &= \mathrm{Dir}(\theta|\alpha) && \text{prior} \\
p(\mathcal{D}|\theta) &= \mathrm{Mu}(x|n, \theta) && \text{likelihood}
\end{aligned}
$$

Infer

$$
p(\theta|\mathcal{D}) = \mathrm{Dir}(\theta|\alpha + x) \qquad \text{posterior}
$$

Probabilistic inference: general recipe

Story

Learn something ...

Specify

- Prior
- Likelihood

Infer

- Posterior
- MAP, Posterior predictive distribution

Why MAP is sometimes dangerous

part 1: Transformation of variables

Note:

- On the following slides we are using small letters for random variables, since we are talking about transformations...
- This way it is less ugly, and less confusing (?).
- Sorry!

Transformation of variables (1)

Theorem 7.1 (transformation of variable)

Suppose $y(x)$ is an increasing monotonic function of some random variable x with PDF $p_x(x)$.

1. *Since $y(x)$ is a monotonic function, it is invertible, i.e. also x can be seen as a function $x(y)$.*
2. *y is also a random variable.*
3. *The PDF $p_y(y)$ is as follows related to $p_x(x)$:*

$$p_y(y) = p_x(x(y))\frac{dx(y)}{dy}$$

Informal proof: preserve probability mass $p_x(x)dx = p_y(y)dy$.

Note: we omit the absolute values around dx/dy since we assume that the transformation is increasing.

Example: x with PDF $p_x(x)$, $y = \log x$. Then $p_y(y) = p_x(\exp(y))\exp(y)$.

Transformation of variables (2)

Informal formula to remember:

$$p(x)dx = p(y)dy$$

Theorem 7.2 (rule of the lazy statistician)

Given a random variable x with PDF $p(x)$ the expected value of $y(x)$ is

$$E(y) = \int y(x)p(x)dx$$

This rule is lazy, because there is no need to find $p(y)$.

From Wasserman, All of Statistics, Theorem 3.6.

Why MAP is sometimes dangerous

part 2: Example

Extended transformation example (1)

Beta distribution:

$$p(\pi) = \mathrm{Beta}(\pi|a, b) = \frac{1}{B(a, b)}\pi^{a-1}(1-\pi)^{b-1} \text{ for } \pi \in [0, 1]$$

Transformation:

$$x(\pi) = \log\frac{\pi}{1-\pi} \text{ and its (well-known) inverse } \pi(x) = \frac{1}{1+e^{-x}}$$

What is $p(x)$?

Answer:

$$\begin{aligned} p(x) &= \mathrm{Beta}(\pi(x)|a, b)\frac{d\pi}{dx} \\ &= \frac{1}{B(a, b)}\pi(x)^{a-1}(1-\pi(x))^{b-1}\pi(x)(1-\pi(x)) \\ &= \frac{1}{B(a, b)}\pi(x)^{a}(1-\pi(x))^{b} \end{aligned}$$

Extended transformation example (2)

Mean with and w/o transformation:

$$E(\pi) = \frac{a}{a+b} \qquad = \int \pi\, p(\pi)\, d\pi$$

$$E(x) = \log\frac{a}{b} = x(E(\pi)) \qquad = \int x\, p(x)\, dx$$

Mode with and w/o transformation, i.e. maximum of PDF:

$$\arg\max_{\pi} p(\pi) = \frac{a-1}{a+b-2} \text{ for } a, b > 1$$

$$\arg\max_{x} p(x) = \log\frac{a}{b} \neq x\left(\frac{a-1}{a+b-2}\right)$$

DANGER!

DANGER:

- Mean doesn't change under transformation (define as integral).
- Mode/maximum might change after transformation!
- So be careful with maximum a posteriori (MAP) estimates...

Naming conventions

- *MAP* is “maximum a-posteriori”.
- The *MAP estimator* for a parameter θ is a function of observed data, that calculates the value for θ, that maximizes the posterior distribution.
- *ML* is “maximum likelihood”.
- The *ML estimator* (sometimes called *MLE*) for a parameter θ is a function of observed data, that calculates the value for θ, that maximizes the likelihood.

MAP vs ML

$$
\begin{aligned}
\theta_{\text{MAP}} &= \arg\max_{\theta} p(\theta|\mathcal{D}) \\
&= \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) && \text{"Bayes rule"} \\
&= \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta) && \text{"}p(\mathcal{D})\text{ is const wrt }\theta\text{"} \\
&= \arg\max_{\theta} \log p(\mathcal{D}|\theta) + \log p(\theta) && \text{"log is monotone"} \\
&= \arg\min_{\theta} -\log p(\mathcal{D}|\theta) - \underbrace{\log p(\theta)}_{\text{regularization}}
\end{aligned}
$$

$$
\begin{aligned}
\theta_{\text{ML}} &= \arg\max_{\theta} p(\mathcal{D}|\theta) \\
&= \arg\max_{\theta} \log p(\mathcal{D}|\theta) \\
&= \arg\min_{\theta} \underbrace{-\log p(\mathcal{D}|\theta)}_{\text{negative log-likelihood}}
\end{aligned}
$$

MAP vs ML

Example: Estimate the mean of a Gaussian distribution after seeing data $x_1, x_2, \ldots, x_n$ (just real numbers, univariate):

$$\begin{aligned}
\mu_{\text{MAP}} &= \arg\min_{\mu} -\log p(\mathcal{D}|\mu) - \log p(\mu) \\
&= \arg\min_{\mu} \underbrace{\sum_{i=1}^{n} (x_i - \mu)^2}_{\text{least squares}} + \underbrace{\lambda \|\mu\|^2}_{\text{regularization}} = \frac{1}{n+\lambda} \sum_{i=1}^{n} x_i
\end{aligned}$$

$$\begin{aligned}
\mu_{\text{ML}} &= \arg\min_{\mu} -\log p(\mathcal{D}|\mu) \\
&= \arg\min_{\mu} \underbrace{\sum_{i=1}^{n} (x_i - \mu)^2}_{\text{negative log-likelihood}} = \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}$$

Nice interpretation of MAP

Example: Estimate the mean of a Gaussian distribution after seeing data $x_1, x_2, \ldots, x_n$ (just real numbers, univariate):

$$\mu_{\text{MAP}} = \frac{1}{n+\lambda} \sum_{i=1}^{n} x_i$$

- E.g. $\lambda = 1$ is like adding another (older) observation $x_0 = 0$ and doing ML.
- E.g. $\lambda = 2$ is like adding two (older) observations with value zero and doing ML.
- E.g. $\lambda = 100$ is like adding 100 (older observations with value zero and doing ML.

Notes:

- The MLE is like MAP with $\lambda = 0$, i.e. without previous observations.
- For an integer λ we can interpret the MAP estimator as an MLE with λ many additional zero measurements.
- The similarity to the parameters a and b of the Beta distribution which can also be interpreted as previous observations.