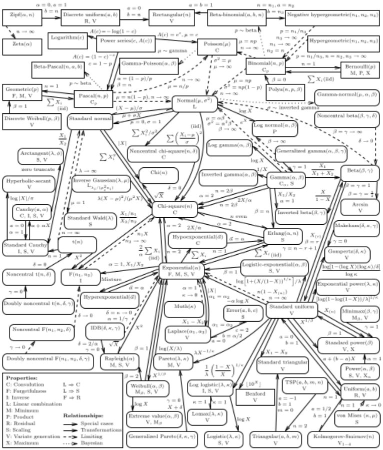
Machine Learning

Lecture 5

Last time:

Statistical model = parametrized family of probability distributions



Last time:

Statistical model = parametrized family of probability distributions

„Point estimation":

$$\left.\begin{array}{c}\text{Data}\\ +\\ \text{Statistical model}\end{array}\right\} \longrightarrow \begin{array}{c}\text{a single}\\ \text{distribution}\end{array}$$

Data $x_1, \dots, x_n$

Examples:

- Maximum Likelihood estimation

$$\hat{\theta}_{ML} := \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} p(x_i \mid \theta)$$

$\operatorname*{argmax}_{\theta}$

$\underbrace{p(D \mid \theta)}_{\text{likelihood function}}$

really $\operatorname*{argmax}_{x} f(x)$ is a <u>set</u>

but we just assume that we can pick an element.

Examples:

Data $x_1, \dots, x_n$

- Maximum Likelihood estimation

$$\hat{\theta}_{ML} := \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} p(x_i \mid \theta)$$

- Maximum a posteriori estimation

posterior

likelihood

prior

$$\hat{\theta}_{MAP} := \operatorname*{argmax}_{\theta} \boxed{p(\theta \mid x_1, \dots, x_n)}$$

$$= \operatorname*{argmax}_{\theta} \frac{\boxed{\prod_{i=1}^{n} p(x_i \mid \theta)} \cdot p(\theta)}{\prod_{i=1}^{n} p(x_i)}$$

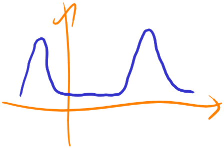
Mode of posterior distr.

$$= \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} p(x_i \mid \theta) \cdot p(\theta)$$

$$= \operatorname*{argmax} \log \prod_{i=1}^{n} p(x_i \mid \theta) \cdot p(\theta) = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} p(x_i \mid \theta) p(\theta)$$

mode of a distribution: $\underset{x}{\arg\max}\ \underline{p(x)}$

prob. mass function
or
density function



---

mean (= expectation) of a distr.: $E(\theta)$ ...

---

median: (if $\theta \in \mathbb{R}$)

The $m \in \mathbb{R}$ s.t.

$$P(\theta \geq m) = 50\%$$
$$P(\theta \leq m) = 50\%$$

What are good choices of priors?

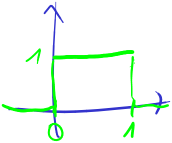
E.g.

- incorporating domain knowledge
- uninformative priors
- from a conjugate family

e.g. coin flip, Bernoulli $P(X=1)=p$

Ber(p) $P(X=0)=1-p$

$p \in [0,1]$



uninformative prior: uniform distr. on $[0,1]$

Conjugate priors: $p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{p(D)}$

Definition *Let X be a random variable with distribution $X \sim f(x \mid \theta)$ ($\theta \in \Theta$ unknown). A collection $\mathcal{C}$ of probability density functions, or probability mass functions, is called a* conjugate prior family *for the family $\{f(x \mid \theta) \mid \theta \in \Theta\}$, if, whenever one chooses a prior from $\mathcal{C}$, the posterior is also from $\mathcal{C}$.*

$P(X=k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$

Example: Binomial family as statistical model

Beta distr.: $B(a,b)(\theta) := C \cdot \theta^{a-1}(1-\theta)^{b-1}$ $(\theta \in [0,1])$

$$C := \left(\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,d\theta\right)^{-1}$$

$=$ the constant assuring that $\int_0^1 B(a,b)(\theta)\,d\theta = 1$.

Say we have as data D: $\begin{cases} n & \text{coin throws} \\ k & \text{heads.} \end{cases}$

$$p(\theta|D) = \frac{p(D|\theta)\cdot p(\theta)}{p(D)} \leftarrow \text{Beta}(a,b)(\theta)$$

$$\propto \underbrace{\binom{n}{k}\theta^k(1-\theta)^{n-k}}_{p(D|\theta)} \cdot \theta^{a-1}\cdot(1-\theta)^{b-1}$$

$$\propto \theta^{k+a-1}(1-\theta)^{n-k+b-1}$$

$$\propto \text{Beta}(k+a,\, n-k+b)(\theta)$$

$B(1,1)$ is an uninformative prior: [corresponds to assuming both outcomes are equally likely]

$\|$

uniform distr. on $[0,1]$.

E.g. for the beta-binomial model

- MAP and ML

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

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

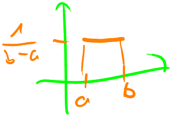
$$
\begin{aligned}
p(x = 1 \mid \mathcal{D}) &= \int_0^1 p(x = 1 \mid \theta) p(\theta \mid \mathcal{D}) d\theta \\
&= \int_0^1 \theta \, \text{Beta}(\theta \mid a+k, b+n-k) d\theta \\
&= \frac{a+k}{a+b+n} = \text{posterior mean}
\end{aligned}
$$

Some families & conjugate families:

Distribution	parameter	conjugate family
Bernoulli(p)	p	Beta
Binomial(k,p)	p	Beta
Geometric(p)	p	Beta
Poisson(λ)	λ	Gamma
Multinomial($p_1, \ldots, p_n$)	$p_1, \ldots, p_n$	Dirichlet
univariate Normal(μ, σ^2)	μ	Normal
univariate Normal(μ, σ^2)	σ^2	Inverse Gamma
multivariate Normal($\boldsymbol{\mu}, \boldsymbol{\Sigma}$)	$\boldsymbol{\mu}$	multivariate Normal
multivariate Normal($\boldsymbol{\mu}, \boldsymbol{\Sigma}$)	$\boldsymbol{\Sigma}$	inverse Wishart
Uniform($[0, \theta]$)	θ	Pareto

ML vs. MAP

- Often, for uniform prior: $\hat{\theta}_{ML} = \hat{\theta}_{MAP}$



$$p(\theta|D) = \frac{p(D|\theta)\cdot p(\theta)}{p(D)} = \frac{p(D|\theta)\,\frac{1}{b-a}}{p(D)} \propto \underbrace{p(D|\theta)}_{\text{likelihood function}}$$

- In general: Prior introduces a bias to $\hat{\theta}_{ML}$ towards mode of the prior.

- For lots of data: $\hat{\theta}_{MAP}$ converges to $\hat{\theta}_{ML}$

$$\hat{\theta}_{ML} = \arg\max_{\theta} \log p(\mathcal{D}|\theta) = \arg\max \sum \log p(D|\theta)$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(\mathcal{D}|\theta) + \log p(\theta) = \arg\max \sum \log p(D|\theta) + \log p(\theta)$$

Many summands $\longrightarrow$ $\log p(\theta)$ unimportant.

MAP estimator and ML estimator

- let's denote the data as $\mathcal{D}$ (was k in the beta-binomial model)
- summarize the posterior by a point estimate
- maximum a posteriori estimator (MAP)

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

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

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

- 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

ML vs MAP: insights

ML is minimizing the negative log-likelihood:

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

MAP is a regularized ML:

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

Famous ML estimator for Gaussian likelihoods

Setup

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

Maximize the likelihood:

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

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

ML vs MAP: comparing the estimators

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

$$p(\mu) = \mathcal{N}(\mu \,|\, 0, \tau^2) \qquad \text{prior mean}$$

$$p(x_1, \dots, x_n \,|\, \mu) = \prod_{i=1}^{n} \mathcal{N}(x_i \,|\, \mu, \sigma^2) \qquad \text{likelihood of the data}$$

For where $\lambda = \sigma^2/\tau^2$ we can derive:

$$\begin{aligned}
\mu_{\text{MAP}} &= \arg\min_{\mu} -\log p(x_1, \dots, x_n \,|\, \mu) - \log p(\mu) = \dots \\
&= \arg\min_{\mu} \underbrace{\sum_{i=1}^{n} (x_i - \mu)^2}_{\text{least squares}} + \underbrace{\lambda \mu^2}_{\text{regularization}} = \frac{1}{n + \lambda} \sum_{i=1}^{n} x_i \\
\mu_{\text{ML}} &= \arg\min_{\mu} -\log p(\mathcal{D} \,|\, \mu) \\
&= \arg\min_{\mu} \underbrace{\sum_{i=1}^{n} (x_i - \mu)^2}_{\text{negative log-likelihood}} = \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}$$

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):

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

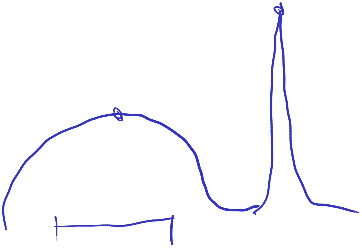
- E.g. $\lambda = 1$ (i.e. $\sigma^2 = \tau^2$) is like adding another (older) observation $x_0 = 0$ and doing ML.
- E.g. $\lambda = 2$ (i.e. $\sigma^2 = 2\tau^2$) is like adding two (older) observations with value zero and doing ML.
- E.g. $\lambda = 100$ (i.e. $\sigma^2 = 100\tau^2$) is like adding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

Notes:

- The MLE is like MAP with $\lambda = 0$ (i.e. $\tau^2 = \infty$, thus having an infinitely wide Gaussian prior), i.e. without previous observations.
- For any integer λ we can interpret the MAP estimator as an MLE with λ many additional zero measurements.
- Parameter λ is similar to parameters a and b of the Beta distribution which also count previous observations.

Other estimators?

Can we use the whole posterior distribution?



Which estimator 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 $p(\theta \mid \mathcal{D})$ pick the $\hat{\theta}$ that minimizes the *posterior expected loss*

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

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

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

Which estimator 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 the MAP estimator

- 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

$\overline{0\text{-}1}$-loss gives MAP in the discrete case:

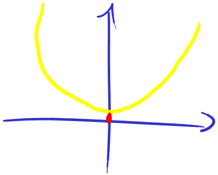
$$\rho(\hat{\theta}) = \int L(\hat{\theta},\theta)p(\theta\,|\,\mathcal{D})\,d\theta \qquad L(\hat{\theta},\theta) = \begin{cases} 0 & \hat{\theta} \neq \theta \\ -1 & \hat{\theta} = \theta \end{cases}$$

(discrete) $$= \sum_{\theta} L(\hat{\theta},\theta)p(\theta|D)$$

$$= -p(\hat{\theta}|D) \qquad \text{(other summands are } = 0\text{)}$$

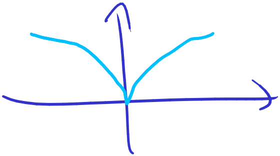
[0-1-loss from previous slide: $\ldots = 1 - p(\hat{\theta}|D)$]

Quadratic loss: $L(\hat\theta,\theta) = (\theta-\hat\theta)^2$



Expected loss.

$$\rho(\hat\theta) = E((\theta-\hat\theta)^2)$$
$$= E(\theta^2 - 2\theta\hat\theta + \hat\theta^2)$$
$$= E(\theta^2) - 2E(\theta)\cdot\hat\theta + (\hat\theta)^2$$



$(\hat\theta-\theta)^{0.2}$

Minimize:

$$0 \overset{!}{=} \frac{\partial}{\partial\hat\theta} = -2E(\theta) + 2\hat\theta$$

$$\Rightarrow \hat\theta = E(\theta)$$ posterior expectation

Which estimator should I choose? (3)

Story:

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

What loss function should you use? What is the resulting estimator?
(Here you should use l_1 loss to minimize the distance to the elevator...)

Summary of point estimators

- Maximum Likelihood estimator (MLE):

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

- Bayes estimator:
  - Maximum Aposteriori (MAP) estimator (minimizes 0-1 loss):

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta} p(\theta \,|\, \mathcal{D}) \\ &= \arg\max_{\theta} p(\mathcal{D} \,|\, \theta) p(\theta)\end{aligned}$$

"the mode of the posterior

  - Posterior mean (the estimator minimizing quadratic loss):

$$\theta_{\text{posterior mean}} = \mathsf{E}_{\theta}\, p(\theta \,|\, \mathcal{D})$$

  - Posterior median (the estimator minimizing l_1 loss):

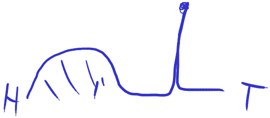
$$\theta_{\text{posterior median}} = \ldots$$

i.e. $\int_{\theta<\theta_{\text{posterior median}}} p(\theta \,|\, \mathcal{D}) d\theta = \int_{\theta>\theta_{\text{posterior median}}} p(\theta \,|\, \mathcal{D}) d\theta$

What else can we do with the posteriors?

Don’t we usually just want point estimates?

Posterior predictive distribution



Alternative to point estimates such as ML and MAP:

- posterior expresses our belief state about the world, e.g.

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

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

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

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

E.g. for the beta-binomial model

- MAP and ML

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

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

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

Inference for a difference in proportions

MLPP 5.2.3, see link in MLPP for the source

θ = degree of reliability

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)

$$p(\theta_1 \mid \mathcal{D}_1) = \mathrm{Beta}(\theta_1 \mid 91, 11) \qquad \text{posterior about reliability}$$

$$p(\theta_2 \mid \mathcal{D}_2) = \mathrm{Beta}(\theta_2 \mid 3, 1) \qquad \text{posterior about reliability}$$

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

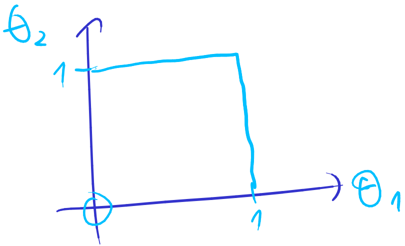
$$p(\theta_1 > \theta_2 \mid \mathcal{D}_1, \mathcal{D}_2)$$

$$= \int_0^1 \int_0^1 [\theta_1 > \theta_2]\, \mathrm{Beta}(\theta_1 \mid 91, 11)\, \mathrm{Beta}(\theta_2 \mid 3, 1)\, d\theta_1 d\theta_2 \approx 0.710$$

using numerical integration (your exercise...).

$$\underset{(\theta_1,\theta_2)}{[\theta_1 > \theta_2]} := \begin{cases} 1 & \text{if } \theta_1 > \theta_2 \\ 0 & \text{else} \end{cases}$$

Θ_1, Θ_2 joint distr.



Monte Carlo simulation:

Draw e.g. 1000 random samples (Θ_1, Θ_2)

& compute the fraction where $\Theta_1 > \Theta_2$.