Lecture 3 - Bayesian Parameter Estimation

CS5487 Lecture Notes (2022B)
Prof. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Problem MLE

Coin Bernoulli R.V. $= \{0 = T, 1 = H\}$

MLE: $\hat{\pi} = \frac{1}{N}\sum_i x_i$

Suppose we see: $D = \{1,1,1,0,0,0,0\} \Rightarrow \hat{\pi} = \frac{3}{7}$ ✓

What if we see $D = \{1,1,1\}$ only $\Rightarrow \hat{\pi} = \frac{3}{3} = 1$?

This is unreasonable! we can only see H from this coin. (we never see tails!)

This is an example of overfitting (not enough samples to get a good estimate of the parameter.)

- use our knowledge: we know $\pi \approx \frac{1}{2}$ for most coins. Incorporate this knowledge into our estimate of π.

Bayesian Param Estimation

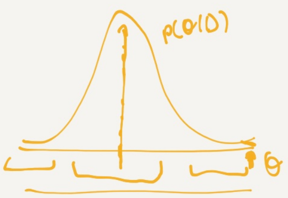
- treat θ as a r.v.
- Framework
  - training set $D = \{x_1, \ldots, x_N\}$
  - prob. density given parameter θ: $p(x_i|\theta)$
  - prior distribution on parameter θ: $p(\theta)$
    (encodes prior beliefs about θ, e.g. $\pi \approx \frac{1}{2}$)
- posterior dist. of θ given data D:

$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{\int p(D|\theta)\,p(\theta)\,d\theta} \quad \text{(Bayes' Rule)}$$

- predictive dist. – likelihood of new x_* given data D,

$$p(x_*|D) = \underbrace{\int p(x_*|\theta)\,p(\theta|D)\,d\theta}_{\text{average over all } \theta \text{, weighted by posterior } p(\theta|D)}$$

"allow different explanations of the data"



Example: Gaussian (known variance)

prior on μ: $p(\mu) = N(\mu | \mu_0, \sigma_0^2)$ known

likelihood of x: $p(x|\mu) = N(x|\mu, \sigma^2)$ known

Dataset: $D = \{x_1, \ldots, x_N\}$

Calculate the posterior

$$p(\mu | D) = \frac{\left[\prod_i p(x_i|\mu)\right] p(\mu)}{\int \left[\prod_i p(x_i|\mu)\right] p(\mu)\, d\mu}$$

← product of Gaussians.

← doesn't depend on μ. (constant wrt μ)

☆ Just look at numerator wrt μ, then normalize later.

first 2 terms:

$$p(x_1|\mu)p(x_2|\mu) = N(x_1|\mu, \sigma^2) N(x_2|\mu, \sigma^2)$$

$$= N(\underset{\text{"x"}}{\mu} | \underset{a}{x_1}, \underset{A}{\sigma^2})\, N(\underset{\text{"x"}}{\mu} | \underset{b}{x_2}, \underset{B}{\sigma^2})$$

$$= \cancel{N}(x_1 | x_2, 2\sigma^2) N(\mu | \tilde{\mu}_2, \tilde{\sigma}_2^2)$$

$$\begin{cases} \dfrac{1}{\tilde{\sigma}_2^2} = \dfrac{1}{\sigma^2} + \dfrac{1}{\sigma^2} = \dfrac{2}{\sigma^2} \\ \tilde{\mu}_2 = \dfrac{\sigma^2}{2}\left(\dfrac{x_1}{\sigma^2} + \dfrac{x_2}{\sigma^2}\right) = \dfrac{1}{2}(x_1 + x_2) \end{cases}$$

$$p(x_1|\mu)p(x_2|\mu) \propto N(\mu | \tilde{\mu}_2, \tilde{\sigma}_2^2)$$ (throw away the constant factor)

product of Gaussians (PS1-7)

$$N(x|a,A)N(x|b,B) = N(a|b, A+B)\, N(x|c, C)$$

$$C = \frac{1}{1/A + 1/B} \Rightarrow \frac{1}{C} = \frac{1}{A} + \frac{1}{B}$$

$$c = C\left(\frac{a}{A} + \frac{b}{B}\right)$$

$(x-\mu)^2 = (\mu - x)^2$

first 3 terms

$$N(\mu | \tilde{\mu}_2, \tilde{\sigma}_2^2) N(x_3 | \mu, \sigma^2) \propto N(\mu | \tilde{\mu}_3, \tilde{\sigma}_3^2)$$

$$\underbrace{\frac{1}{\tilde{\sigma}_3^2}}_{\text{precision}} = \frac{1}{\tilde{\sigma}_2^2} + \frac{1}{\sigma^2} = \frac{2}{\sigma^2} + \frac{1}{\sigma^2} = \frac{3}{\sigma^2}$$

$$\tilde{\mu}_3 = \frac{\cancel{\sigma^2}}{3}\left(\frac{\frac{1}{\cancel{2}}(x_1+x_2)}{\frac{\cancel{\sigma^2}}{\cancel{2}}} + \frac{x_3}{\cancel{\sigma^2}}\right) = \frac{1}{3}(x_1 + x_2 + x_3)$$

⋮

first N terms:

$$\prod_i p(x_i|\mu) \propto N(\mu | \tilde{\mu}_N, \tilde{\sigma}_N^2) \begin{cases} \tilde{\mu}_N = \frac{1}{N}\sum_i x_i = \hat{\mu}_{ML} \\ \tilde{\sigma}_N^2 = \sigma^2/N \end{cases}$$

x prior

$$N(\mu | \hat{\mu}_{ML}, \tilde{\sigma}_N^2) N(\mu | \mu_0, \sigma_0^2) \propto N(\mu | \hat{\mu}_N, \hat{\sigma}_N^2)$$

$$\frac{1}{\hat{\sigma}_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \Rightarrow \hat{\sigma}_N^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}$$

$$\hat{\mu}_N = \frac{1}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}\left(\frac{\hat{\mu}_{ML}}{\frac{\sigma^2}{N}} + \frac{\mu_0}{\sigma_0^2}\right) \quad \frac{\sigma_0^2\sigma^2}{\sigma_0^2\sigma^2}$$

$$\hat{\mu}_N = \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2}\hat{\mu}_{ML} + \frac{\sigma^2}{\sigma^2 + N\sigma_0^2}\mu_0$$

Finally

$$\boxed{p(\mu | D) = N(\mu | \hat{\mu}_N, \hat{\sigma}_N^2)}$$

What does it mean?

$$\frac{1}{\hat{\sigma}_n^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\hat{\mu}_n = \underbrace{\frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2}}_{\alpha} \hat{\mu}_{ML} + \underbrace{\frac{\sigma^2}{\sigma^2 + N\sigma_0^2}}_{1-\alpha} \mu_0$$

interpolating btwn MLE soln & prior μ_0.

Dataset size

$N = 0 \Rightarrow \alpha = 0 \Rightarrow \hat{\mu}_n = \mu_0$ ← no data, use prior.

$N \to \infty \Rightarrow \alpha = 1 \Rightarrow \hat{\mu}_n = \hat{\mu}_{ML}$ ← lots of data, use MLE

variance

$N = 0 \Rightarrow \hat{\sigma}_n^2 = \sigma_0^2$ ← prior uncertainty

$N \to \infty \Rightarrow \hat{\sigma}_n^2 \to 0$ ← converges to a single value.

$\sigma_0^2 \ll \sigma^2 \Rightarrow \alpha = 0 \Rightarrow \hat{\mu}_n = \mu_0$ (strong belief compared to noise → use our belief.)

$\sigma_0^2 \gg \sigma^2 \Rightarrow \alpha = 1 \Rightarrow \hat{\mu}_N = \hat{\mu}_{ML}$ (weak belief → use MLE)

$$\sigma^2 = \sigma_0^2 \Rightarrow \alpha = \frac{N}{N+1} \Rightarrow \hat{\mu}_n = \frac{N}{N+1}\hat{\mu}_{ML} + \frac{1}{N+1}\mu_0$$

$$= \frac{1}{N+1}\left(N\hat{\mu}_{ML} + \mu_0\right)$$

$$= \underbrace{\frac{1}{N+1}\left(\sum_i x_i + \mu_0\right)}$$

"add a virtual sample at μ_0, then "compute mean"

- for large N, the v.s. doesn't matter.
- for small N, moves the posterior towards μ_0.

[This is a form of regularization]

Predictive distribution

$$p(\mu|D) = N(\mu|\hat{\mu}_n, \hat{\sigma}_n^2)$$

$$p(x|\mu) = N(x|\mu, \sigma^2)$$

$$p(x|D) = \int p(x|\mu)p(\mu|D)\,d\mu$$

$$= \int N(\mu|x, \sigma^2)N(\mu|\hat{\mu}_n, \hat{\sigma}_n^2)\,d\mu$$

$$= \int \underbrace{N(x|\hat{\mu}_N, \sigma^2 + \hat{\sigma}_n^2)}_{\text{no } \mu} \underbrace{N(\mu|\ldots, \ldots)}\,d\mu$$

$$\boxed{p(x|D) = N(x|\hat{\mu}_N, \hat{\sigma}_n^2 + \sigma^2)}$$

same mean as posterior

variance of parameter $\mu | D$ (uncertainty)

uncertainty due to noisy obs.

$\sigma_0^2 \ll \sigma^2$: $1 - \alpha = \frac{\sigma^2}{\sigma^2 + N\sigma_0^2} = 1$

$\Rightarrow \alpha = 0$

Maximum a Posteriori (MAP)

Avoid calculating the denominator of Bayes' Rule:

$$\int p(D|\theta)\,p(\theta)\,d\theta \quad \dots \text{ difficult for many cases.}$$

Soln: pick the θ w/ largest posterior probability.

$$\hat{\theta}_{MAP} = \arg\max_{\theta}\, p(\theta|D)$$

Bayes' Rule

$$= \arg\max_{\theta} \frac{\overset{\text{data like.}}{p(D|\theta)}\,\overset{\text{prior}}{p(\theta)}}{\int p(D|\theta)\,p(\theta)\,d\theta}$$

constant wrt θ (not a function of θ)

$$= \arg\max_{\theta}\, p(D|\theta)\,p(\theta)$$

$$\boxed{\hat{\theta}_{MAP} = \arg\max_{\theta}\, \underbrace{\log p(D|\theta)}_{\text{data LL for MLE}} + \underbrace{\log p(\theta)}_{\text{regularization.}}}$$

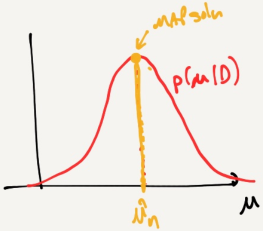
Example: Gaussian

$$\hat{\mu}_{MAP} = \arg\max_{\mu}\, p(\mu|D) = \arg\max_{\mu}\, N(\mu|\hat{\mu}_n, \hat{\sigma}_n^2)$$

$$\hat{\mu}_{MAP} = \hat{\mu}_n$$

Approximate posterior as a delta function:

$$p(\mu|D) \approx \delta(\mu - \hat{\mu}_n)$$



$$p(x|D) \approx p(x|\hat{\mu}_n) = N(x|\hat{\mu}_n, \sigma^2)$$

Bayesian Regression

Same setup as before:

$x \in \mathbb{R}$

$f(x) = \phi(x)^T\theta \leftarrow \theta \in \mathbb{R}^d$

$y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$

$$p(y|x,\theta) = N(y|f(x), \sigma^2)$$

Introduce prior on θ:

$$p(\theta) = N(\theta \,|\, \underset{\mathbb{R}^d}{\underset{\uparrow}{}} \underset{\text{zero mean vector}}{0}, \; \underset{\text{scaled identity covariance matrix}}{\alpha \overset{\text{identity matrix}}{I}})$$

MAP estimate

$$\hat{\theta} = \arg\max_{\theta}\, \log p(D|\theta) + \log p(\theta)$$

$$= \arg\max_{\theta} \sum_i \log p(y_i|x_i, \theta) + \log p(\theta)$$

$$\vdots$$

$$= \arg\min_{\theta}\, \|y - \Phi^T\theta\|^2 + \underset{\text{constant}}{\lambda}\|\theta\|^2$$

constant

controls regularization

$\lambda = 0 \Rightarrow$ LS

ridge regression
- regularized LS
- Tikhonov regularization
- shrinkage
- weight decay

$$\boxed{\hat{\theta} = (\underbrace{\Phi\Phi^T}_{} + \lambda I)^{-1}\Phi y}$$

$\lambda = \frac{\sigma^2}{\alpha}$ (see tutorial)

regularize the covariance matrix to prevent inverting an ill-conditioned matrix

(adds λ to all the eigenvalues of $\Phi\Phi^T$)