Lecture 3: Bayesian Parameter Estimation

Problem w/ MLE

Model a coin as a Bernoulli r.v. $\{0 = T, 1 = H\}$

MLE: $\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Suppose we see: $D = \{1,1,1,0,0,0,1\} \Rightarrow \hat{\pi} = \frac{4}{7}$ ✓

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

$\Rightarrow$ This unreasonable $p(x=0) = 0$ $\Rightarrow$ tails never happens

$\Rightarrow$ an example overfitting (too little data)

How to incorporate our knowledge about coins into our estimate of π? e.g. $\pi = \frac{1}{2}$ typically.

CS5487 Lecture Notes (2024A)
Prof. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Bayesian Parameter Estimation

- treat θ parameter as a r.v.
- Framework
  - training set $D = \{x_1, \ldots, x_n\}$
  - prob function: $p(x_i|\theta)$
  - prior distribution on θ: $p(\theta)$ ⇐
    (encode our beliefs about θ)

- posterior distribution of θ given D

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

- predictive distribution – the likelihood of a new sample x_* given the data D.

$$p(x_*|D) = \underbrace{\int \underbrace{p(x_*|\theta)}_{\text{likelihood given a }\theta}\underbrace{p(\theta|D)}_{\text{posterior of }\theta} d\theta}_{\text{"average over all }\theta\text{ weighted by its posterior probability"}}$$

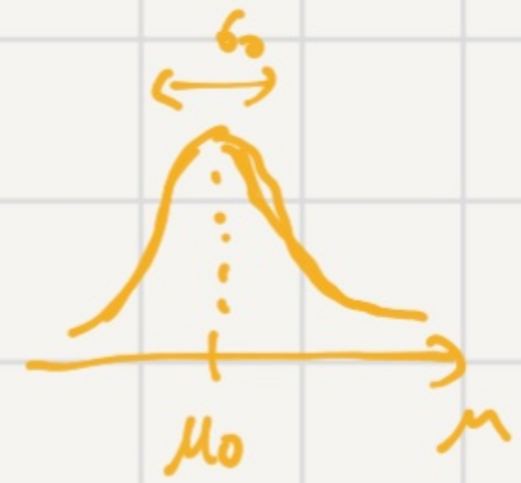
"allow different explanations of the data in terms of θ"

Example Gaussian (w/ known variance)

Dataset: $D = \{x_1, \dots, x_n\}$

likelihood: $p(x_i|\mu) = N(x_i|\underbrace{\mu}_{\text{param}}, \underbrace{\sigma^2}_{\text{fixed}})$

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



Posterior

$$p(\mu|D) = \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)\,d\mu} = \frac{\prod_{i=1}^{n} p(x_i|\mu)\; p(\mu)}{\underbrace{\int \prod_{i=1}^{n} p(x_i|\mu)p(\mu)\,d\mu}_{\text{not a function of } \mu \Rightarrow \text{constant}}}$$

$$= \frac{e^{f(\mu)}}{C}$$ ← what is $f(\mu)$?

Just look at exponent of numerator & normalize later.

$$\log p(\mu|D) \propto \sum_{i=1}^{n} \log p(x_i|\mu) + \log p(\mu)$$

up to an additive constant

$$\propto \sum_{i=1}^{n} \frac{-1}{2\sigma^2}(x_i-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2$$

$$= \frac{-1}{2\sigma^2}\underbrace{\sum_{i=1}^{n}(\cancel{x_i^2}-2\mu x_i+\mu^2)}_{-2\mu\sum_i x_i + n\mu^2} - \frac{1}{2\sigma_0^2}(\mu^2 - 2\mu_0\mu + \cancel{\mu_0^2})$$

$$\sum_i x_i = n\hat{\mu}_{ML}$$

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

$$= -\frac{1}{2}\left[\underbrace{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}_{a}\mu^2 - 2\underbrace{\left(\frac{n}{\sigma^2}\hat{\mu}_{ML}+\frac{1}{\sigma_0^2}\mu_0\right)}_{b}\mu\right]$$

Completing the square

$$ax^2 - 2bx + c = a(x-d)^2 + e$$

$$d = b/a$$

$$e = c - \frac{b^2}{a}$$

$$\Rightarrow \log p(\theta|D) \propto -\frac{1}{2}\underbrace{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}_{\hat{\sigma}_n^{-2}}\left(\mu - \underbrace{\frac{1}{a}\left(\frac{n}{\sigma^2}\hat{\mu}_{ML}+\frac{1}{\sigma_0^2}\mu_0\right)}_{\hat{\mu}_n}\right)^2 + \dots$$

$$= -\frac{1}{2\hat{\sigma}_n^2}(\mu-\hat{\mu}_n)^2 + \text{const}$$

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

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

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

$$\boxed{\hat{\mu}_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_{ML} + \left(\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\right)\mu_0}$$

Thus

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

What does it mean?

$$\hat{\mu}_n = \underbrace{\left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)}_{\alpha} \hat{\mu}_{ML} + \underbrace{\left(\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\right)}_{1-\alpha} \mu_0$$

$\hat{\mu}_{ML} \longleftrightarrow \mu_0$

$$\hat{\sigma}_n^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \quad \infty$$

Datasize n:

$n=0 \Rightarrow \alpha=0 \Rightarrow \hat{\mu}_n = \mu_0$

$n\to\infty \Rightarrow \alpha=1 \Rightarrow \hat{\mu}_n = \hat{\mu}_{ML}$

smoothing btwn our prior belief μ_0 & data-driven $\hat{\mu}_{ML}$

uncertainty

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

$n\to\infty \Rightarrow \hat{\sigma}_n^2 = 0$ ← converges to one value $\hat{\mu}_{ML}$

Belief strength

$\sigma_0^2 \ll \sigma^2 \Rightarrow \alpha=0 \Rightarrow \hat{\mu}_n = \mu_0$ ⇒ use prior

↑ strong belief (small variance)

$\sigma_0^2 \gg \sigma^2 \Rightarrow \alpha=1 \Rightarrow \hat{\mu}_n = \hat{\mu}_{ML}$ ⇒ use MLE

↑ weak belief

$$\sigma_0^2 = \sigma^2 \Rightarrow \alpha = \frac{n\sigma^2}{n\sigma^2+\sigma^2} = \frac{n}{n+1}$$

$$\Rightarrow \hat{\mu}_n = \frac{n}{n+1}\hat{\mu}_{ML} + \frac{1}{n+1}\mu_0$$

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

$$= \frac{1}{n+1}\underbrace{\left(\sum_{i=1}^{n} x_i + \mu_0\right)}_{n+1 \text{ terms in a sum}}$$

"add a virtual sample at μ_0" then compute the mean.

- for large n, the v.s. doesn't matter.
- for small n, moves the MLE towards the prior mean.
- This is a form of <u>regularization</u>.

Predictive Distribution

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

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

predictive:

$$p(x|D) = \int p(x|\mu)\, p(\mu|D)\, d\mu = \int \underbrace{N(x|\mu,\sigma^2)}_{\dots e^{-(x-\mu)^2} = e^{-(\mu-x)^2}} N(\mu|\hat{\mu}_n, \hat{\sigma}_n^2)\, d\mu$$

$$= \int \underbrace{N(\mu|x,\sigma^2)\, N(\mu|\hat{\mu}_n, \hat{\sigma}_n^2)}_{\text{product of 2 Gaussians (PS 1-7)}}\, d\mu$$

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

$\underbrace{N(a|b, A+B)}\,\underbrace{N(x|c,C)}$

$$= \int \underbrace{N(x|\hat{\mu}_n, \sigma^2+\hat{\sigma}_n^2)}\, \underbrace{\cancelto{1}{N(\mu|\dots,\dots)\, d\mu}}$$

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

posterior mean

variance due to noisy observations.

variance of posterior (uncertainty in μ)

Maximum a Posteriori (MAP)

calculating $\int p(D|\theta)p(\theta)d\theta$ is difficult $\Rightarrow$ approximations.

one solution: pick θ w/ highest posterior probability.

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

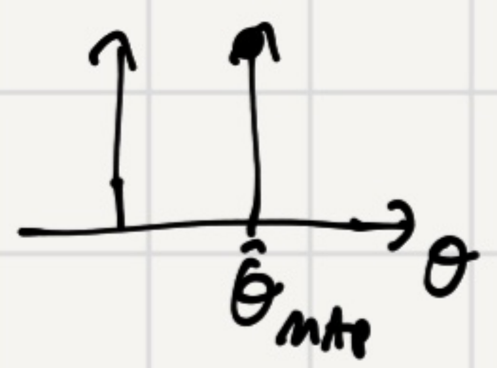
$$= \arg\max_\theta \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \leftarrow \text{constant wrt } \theta$$

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

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

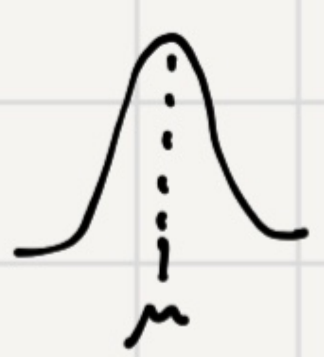
prior belief

posterior: $p(\theta|D) = \delta(\theta - \hat{\theta}_{MAP})$



predictive: $p(x|D) = p(x|\hat{\theta}_{MAP})$

(ok if posterior is peaked, e.g. enough data)



Example: Gaussian

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

$$= \hat{\mu}_n.$$

Bayesian Regression

Similar setup:

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

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

$p(y|x,\theta) = N(y \mid \phi(x)^T\theta, \sigma^2)$

New!

parameter, hyperparameter

$$p(\theta) = N(\theta \mid 0, \alpha I)$$

Consider MAP estimate

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

dropping constants

$$= \underset{\theta}{\arg\max}\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \phi(x_i)^T\theta)^2 - \frac{1}{2\alpha}\|\theta\|^2 \cdot \sigma^2$$

$$= \underset{\theta}{\arg\min}\ \sum_i (y_i - \phi(x_i)^T\theta)^2 + \underbrace{\frac{\sigma^2}{\alpha}}_{\lambda}\|\theta\|^2$$

$$= \underset{\theta}{\arg\min}\ \underbrace{\|y - \Phi^T\theta\|^2}_{\text{squared error (same as L.S.)}} + \lambda \underbrace{\|\theta\|^2}_{\text{controls complexity of } \theta}$$

hyperparameter controls tradeoff.

controls complexity of θ (prevent from taking extreme values)

- regularized L.S.
- ridge regression
- Tikhonov regularization
- Shrinkage
- weight decay

⋮

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

new λI term: regularizes the cov matrix to prevent inverting an ill-conditioned matrix.