Lecture 3 Bayesian Parameter Estimation

Problem of MLE

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

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 head from this coin (we never see tails!)

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

How to deal with this problem?

use our knowledge: we know $\pi \approx \frac{1}{2}$ for most coins

Incorporate this knowledge into our estimate of π

Framework:

- treat θ as a r.v.
- training set $D = \{x_1, \cdots, x_N\}$
- probility distribution given parameter θ: $p(x_i|\theta)$
- prior distribution on parameter θ: $p(\theta)$
  encode prior beliefs about θ. e.g $\pi = \frac{1}{2}$)

Posterior distribution 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 distribution: likelihood of new X_* given data D.

$$p(X_*|D) = \int p(X_*|\theta)\,p(\theta|D)\,d\theta$$

Average over all θ, weighted by posterior $p(\theta|D)$

Allow different explanations of the data.

Example: Gaussian (known variance)

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

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

we know μ_0, σ_0^2, and σ^2

Data set $D = \{x_1, \cdots x_N\}$

Calculate the posterior distribution

$$p(\mu|D) = \frac{\left[\prod_{i=1}^{N} p(D|\mu)\right] p(\mu)}{\int \left[\prod_{i=1}^{N} p(D|\mu)\right] p(\mu)\,d\mu}$$

← product of Gaussian

↓ doesn't depend on μ (consistant w.r.t. μ)

Just look at nomerator w.r.t μ, then normalize later

First two terms:

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

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

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

$= N(\mu|x,\sigma^2)$ without meaning

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

Product of Gaussian:

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

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

$$c = C_1\left(\frac{a}{A}+\frac{b}{B}\right)$$

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

↓ Constant

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

$$p(x_1|\mu)\cdot p(x_2|\mu) \propto N(\mu|\tilde{\mu}_2,\tilde{\sigma}_2^2)$$

Throw away the constant factor

First three items:

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

$$\begin{cases} \frac{1}{\tilde{\sigma}_3^2} = \frac{1}{\tilde{\sigma}_2^2} + \frac{1}{\sigma^2} = \frac{3}{\sigma^2} \\ \tilde{\mu}_3 = \frac{1}{3}\sigma^2\left(\frac{\tilde{\mu}_2}{\tilde{\sigma}_2^2} + \frac{x_3}{\sigma^2}\right) = \frac{\sigma^2}{3}\left(\frac{2}{\sigma^2}\cdot\frac{1}{2}(x_1+x_2) + \frac{x_3}{\sigma^2}\right) \\ \quad = \frac{1}{3}(x_1+x_2+x_3) \end{cases}$$

$\vdots$

First N terms:

$$\prod_{i=1}^{N} p(x_i|\mu) \propto N(\mu|\tilde{\mu}_N, \tilde{\sigma}_N^2)$$

$$\tilde{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x_i = \hat{\mu}_{ML} \qquad \tilde{\sigma}_N^2 = \frac{\sigma^2}{N}$$

$\times$ prior distribution:

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

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

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

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

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

What does it mean ?

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

α $\qquad$ $1-\alpha$

adjust between MLE solution AND 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:

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

$N=0 \Rightarrow \hat{\sigma}_n^2 = \sigma_0^2$ PRIOR UNCERTAINTY

$N \to \infty \Rightarrow \hat{\sigma}_n^2 = 0$ CONVERAGES TO A SINGLE VALUE

$\sigma_0^2 << \sigma^2 \Rightarrow \alpha = 0 \Rightarrow \hat{\mu}_n = \mu_0$

Strongly belief compared to noise $\rightarrow$ USE OUR BELIEF

$\sigma^2 << \sigma_0^2 \Rightarrow \alpha = 1 \Rightarrow \hat{\mu}_n = \mu_{ML}$

weak belief $\rightarrow$ USE MLE

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

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

add a virtual sample at μ_0, then compute mean

Large N: the virtual sample doesn't matter

Small N: moves the posterior towards μ_0

THIS IS A FORM OF REGULARIZATION !

Predictive distribution

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

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

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

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

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

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

$\hat{\mu}_N$: Same mean as posterior

$\hat{\sigma}_n^2$: Variance of Parameter $\mu|D$ (uncertainly)

σ^2 : uncertainly due to noisy observely

Maximum a Posterior (MAP)

Avoid calculate the denominator of Raye's Rule

$\int p(D|\theta)\, p(\theta)\, d\theta$ ---- difficult for many cases

Solution: pick the θ with largest posterior possibility

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

$p(D|\theta)$ → data likelihood

$p(\theta)$ → prior

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

$\int p(D|\theta)\, p(\theta)\, d\theta$: constant w.r.t θ. not a function of θ

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

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

$\log p(D|\theta)$: data Likelihood for MLE

$\log p(\theta)$: Regularization

Example: Gaussian

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

$$= \arg\max_{\mu} N(\mu|\hat{\mu}_n, \hat{\sigma}_n)$$

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

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

$x \in \mathbb{R}$ $\quad f(x) = \phi(x)^T \theta$ $\quad \theta \in \mathbb{R}^d$ $\quad y = f(x) + \varepsilon$ $\quad \varepsilon \sim N(0, \sigma^2)$

Introduce prior on θ: $p(\theta) = N(\theta|0, \alpha I)$

$\theta \in \mathbb{R}^d$

$\theta \in \mathbb{R}^d$

α: scaled identity covariance matrix

MAP estimate

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

$$= \arg\max_\theta \sum_i p(x_i|\theta) + \log p(\theta)$$

$$= \arg\min_\theta \|y - \Phi^T\theta\| + \lambda\|\theta\|^2$$

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

$\lambda = \frac{1}{\alpha}$ controls regularization $\lambda = 0 \Rightarrow$ LS.

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

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