Stochastic Process

Lect 1:

Introduction

Frequentist and bayesian.

Data Matrix call $D = \{x^{(i)}\}_{i=1}^{N}$

$$X = \begin{bmatrix} - x^{(1)} - \\ - x^{(2)} - \\ \vdots \\ - x^{(N)} - \end{bmatrix} \in R^{N \times D}$$

where $x^{(i)} = [x_1^{(i)}, \cdots x_D^{(i)}] \in R^D$

we assume each $x^{(i)}$ is sampled from $P(D|\theta)$ in $\boxed{\text{i.i.d}}$ manner.

Frequentist.: assume θ is a constant.

Then the probability to observe N data points in i.i.d manner is

$$P(D|\theta) = \prod_{i=1}^{N} P(x^{(i)}|\theta)$$

To calculate θ. we can use MLE (Maximum likelihood estimator)

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\; P(\mathcal{D}|\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\; \log P(\mathcal{D}|\theta).$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(x^{(i)}|\theta).$$

Bayesian: Assume θ is not a constant.

$\theta \sim P(\theta) \Leftarrow$ preset prior distribution.

By Baye's Rule: the posterior distribution.

(the prob of θ given the evidence X).

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)\, P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}) \longrightarrow$ Marginal likelihood or evidence.

$$\boxed{= \frac{P(\mathcal{D}|\theta) \cdot P(\theta)}{\int_{\theta} P(\mathcal{D}|\theta) \cdot P(\theta) d\theta}}$$

To get θ, we will maximize the posterior distribution.

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(\theta|\mathcal{D})$$

This is not fully "Bayesian" since $\hat{\theta}_{MAP}$ is a point estimate.

prevent overfitting

$$\stackrel{\Uparrow}{=} \underset{\theta}{\operatorname{argmax}}\, P(\mathcal{D}|\theta)\cdot P(\theta)$$

$\Uparrow$ because $P(\mathcal{D}) = \int_\theta P(\mathcal{D}|\theta)\cdot P(\theta)\,d\theta$ is not a function of θ.

$$= \underset{\theta}{\operatorname{argmax}}\, \log P(\mathcal{D}|\theta) + \log P(\theta).$$

---

posterior distribution. $P(\hat{\theta}_{map}|\mathcal{D})$

$$= \frac{P(\mathcal{D}|\hat{\theta}_{map})\cdot P(\hat{\theta}_{map})}{P(\mathcal{D})}$$

It is a function of $\mathcal{D}$.

likelihood function.

---

predictive uncertainty:

The uncertainty in the prediction induced by $\boxed{\text{uncertainty in the parameter}}$

Compute posterior predictive distribution

$$P(\underline{y} \mid x, D)$$

$$= \int_{\theta} P(y \mid x, \underline{\underline{\theta}})\, P(\underline{\underline{\theta}} \mid D)\, d\theta.$$

↳ Marginalizing out the parameter.
reduce the overfitting//