- Review SRM principle
- Bayesan score / BIC

↑

Model Selection

We know a collection of function classes

$$\mathcal{F}_i = \{ f(\cdot\,; \theta) : \theta \in \mathbb{R}^{d_i},\ i = 1, \dots, K \}$$

Training samples $\mathcal{D}_n = \{ (x_t, y_t) \}_{t=1}^{n}$

Two defn of RISK:

① Expect Risk given parameter θ

$$R(\theta) = \mathop{\mathbb{E}}_{(x,y)\sim P} [\, \text{loss}_{0-1}(y; f(x; \theta)) \,]$$

joint dist.

$$= \int_{X \times \{-1,+1\}} \text{Loss}_{0-1}(y; f(x; \theta))\, d\, P(x, y)$$

If we are working in a particular $f^{\wedge}$ class $\mathcal{F}_i$

$$\hat{\theta}_i = \mathop{\text{argmin}}_{\theta:\, f(\cdot\,;\theta) \in \mathcal{F}_i} J(\theta) \Rightarrow$$ choose the ope $\hat{\theta}_i$ in every function class $\mathcal{F}_i$

$$J(\theta) = \frac{1}{n} \sum_{t=1}^{n} \text{Loss}\big(y_t; f(x_t; \theta)\big) + \lambda \|\theta\|^2$$

↓

Eg. SVM ⇝ hinge loss

Logistic Regression ⇝ logistic loss

Best expected risk within $\mathcal{F}_i$: $R(\theta_i) = R(\hat{f}_i)$, where $\hat{f}_i = f(\cdot\,; \hat{\theta}_i)$

Rmk: • exp. risk can not be computed since $P(x,y)$ is Unknown.

• Approximate the exp. risk with the empirical risk.

② **Emp. Risk → θ**

$$R_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \text{Loss}_{0\text{-}1}(y_t ; f(x_t; \theta))$$

$R_n(\hat{f}_i) = R_n(\hat{\theta}_i)$ $\quad f_i \in \mathcal{F}_i$.

↑ Best empirical risk

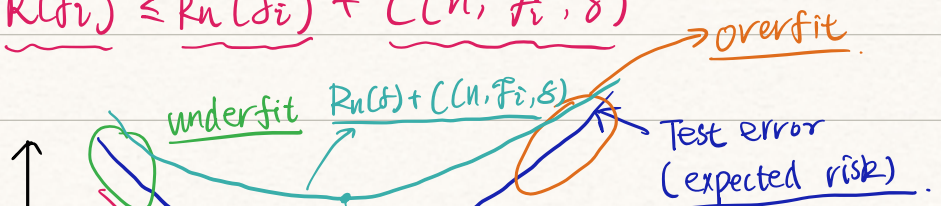
⇓ can be calculated

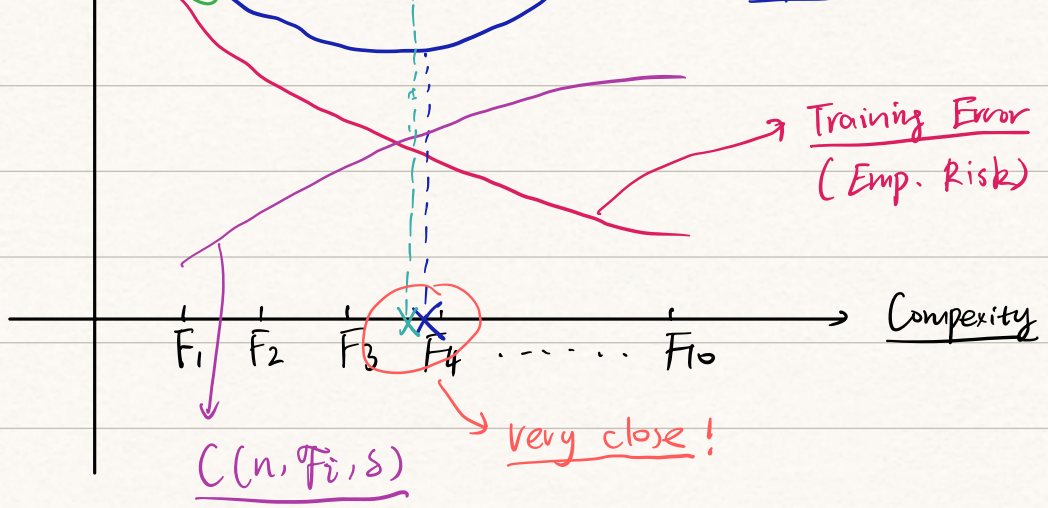
Thm: With probability at least $1-\delta$.

$$\forall f \in \mathcal{F}_i \qquad R(f) \le R_n(f) + C(n, \mathcal{F}_i, \delta)$$

$$C(n, \mathcal{F}_i, \delta) = \sqrt{\frac{\log|\mathcal{F}_i| + \log(\frac{2}{\delta})}{2n}}$$

$$\Rightarrow R(\hat{f}_i) \le R_n(\hat{f}_i) + C(n, \mathcal{F}_i, \delta)$$





→ ERM $\hat{\theta}_i = \underset{\theta : f(\cdot;\theta) \in \mathcal{F}_i}{\operatorname{argmin}} \frac{1}{n} \sum \text{Loss}_{01}(y_t, f(x_t))$

select the most complex model

$i^*_{ERM} = \underset{i=1,\dots K}{\operatorname{argmin}} R_n(f_i) = \boxed{K}$

→ SRM $i^* = \underset{i=1,\dots,K}{\operatorname{argmin}} R_n(f_i) + C(n, \mathcal{F}_i, \delta)$

fit to model

penalize more complex models

Bayesian Score

Suppose that moder $\mathcal{F}$ takes $x \in \mathbb{R}^d$ and maps to $y \in \mathbb{R}$

$$P(y \mid \underline{x}, \underline{\theta}, \sigma^2) = \mathcal{N}(y; \underline{\theta}^T \underline{x}, \sigma^2)$$

Given a datasee $\{(\underline{x}_t, y_t)\}_{t=1}^n \subset \mathbb{R}^d \times \mathbb{R}$. we want to estimate $\underline{\theta}$.

⇓ Method

Maximize Likelihood!

Likelihood : $\mathcal{L}(\mathcal{D}; \theta) = \prod_{t=1}^{n} P(y_t | x_t, \theta, \sigma^2)$

$$= \prod_{t=1}^{n} N(y_t; \theta^T x_t, \sigma^2)$$

$$= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ - \frac{(y_t - \theta^T x_t)^2}{2\sigma^2} \right\}$$

$$X = \begin{bmatrix} x_1^T & 1 \\ \vdots & \\ x_n^T & 1 \end{bmatrix}_{n \times (d+1)} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

$\Rightarrow$ MLE $\hat{\theta} = \operatorname{argmin} \mathcal{L}(\mathcal{D}_n; \theta)$

$= (X^T X)^{-1} X^T y$. (Assume X – full column rank)

Here we have point estimate of $\hat{\theta}$, namely $(X^T X)^{-1} X^T y$

↓

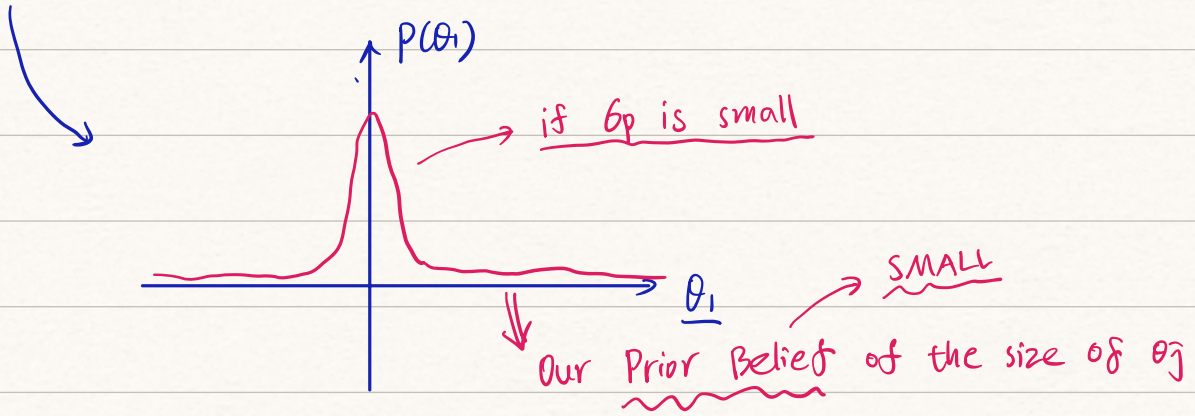
In Bayesian analysis, we want a dist. over θ given $\mathcal{D}_n$

$\mathcal{D} = \{(x_t, y_t)\}_{t=1}^{n} \qquad (x_t, y_t) \sim P$

Method: impose a prior distribution over θ $\rightsquigarrow$ $P(\theta) = N(\theta; 0, \sigma_p^2 I)$

$$P(\theta) = \prod_{j=1}^{d} N(\theta_j; 0, \sigma_p^2) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ - \frac{1}{2\sigma_p^2} \theta_j^2 \right\}.$$

Rmk: σ_p is small, we expect that each component of θ is small



Bayesian Rule (Fundamental relation between Bayesian statistics)

$$\text{Posterior} := \frac{\text{Likelihood} \times \text{Prior}}{\text{Model Evidence}}$$

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

~~Approximately~~ → Equivalent

Rmk: Model evidence ⇔ Bayesian Score (BIC is approximator)

⇔ Marginal likelihood (two names)

equivalent

$$P(\mathcal{D}) = P(\mathcal{D}|\mathcal{F}) = \int L(\mathcal{D};\underline{\theta}')\,P(\underline{\theta}')\,d\underline{\theta}'$$

Model class currently assuming

$$\rightarrow \hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\mathcal{D})$$

depend on the f^ class $\mathcal{F}$

$$= \arg\max_{\theta} \boxed{L(\mathcal{D};\theta)}\,\boxed{P(\theta)}$$

prior (before $\mathcal{D}$)

after $\mathcal{D}$ (observation)

In the linear Gaussian Model

$$\begin{cases} P(\theta) = \mathcal{N}(\theta; 0, \sigma_p^2 I) \\ \mathcal{L}(\theta; \mathcal{D}) = \prod_{t=1}^{n} \mathcal{N}(y_t; x_t^T \theta, \sigma^2) \end{cases}$$

$\Downarrow$

$$\log P(\mathcal{D}|\mathcal{F}) = -\frac{n}{2}\log(2\pi\sigma^2) + \frac{d}{2}\log\lambda - \frac{1}{2}\log|X^TX + \lambda I|$$

$\log P(\mathcal{D})$

$$\lambda = \frac{\sigma^2}{\sigma_p^2}$$

$$- \frac{1}{2\sigma^2}\left(\|y\|^2 - y^T X (X^TX + \lambda I)^{-1} X^T y\right)$$

Actually this is very hard to get

↘ Laplace Approximation

For the linear Gaussian Model prior on θ, we can find the Model Evidence in CLOSED FORM.

↓

Generally, it is impossible for us to calculate the Model Evidence directly!

$\Downarrow$

Laplace Approximation

Linear Gaussian Model

$$P(\theta|\mathcal{D}) = \frac{P(\theta)\mathcal{L}(\mathcal{D}|\theta)}{P(\mathcal{D})} \rightsquigarrow \text{posterior distribution}$$

$$P(\theta|\mathcal{D}) \sim \mathcal{N}(\theta; \mu, \Sigma)$$

$$\begin{cases} \mu = (X^TX + \lambda I)^{-1} X^T y \\ \Sigma = \sigma^2 (X^TX + \lambda I)^{-1} \end{cases}$$

Rmk: $\Sigma = \sigma^2 (X^TX + \lambda I)^{-1}$ (σ^2 → noise variance)

i) depends on the original noise variance

ii) $X^TX = \sum_{t=1}^{n} x_t x_t^T$

$$\Rightarrow \frac{1}{n} X^TX = \frac{1}{n} \sum x_t x_t^T$$

→ If $x_t \sim P(x)$ i.i.d. manner

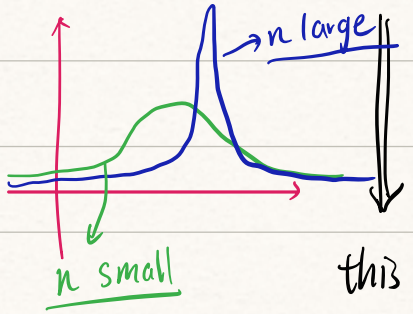
then $\frac{1}{n} \sum x_t x_t^T \xrightarrow{P} C =$

$$\Sigma = \sigma^2 (X^TX + \lambda I)^{-1} = \sigma^2 (n (\tfrac{1}{n} X^TX + \tfrac{1}{n} \lambda I))^{-1}$$

($\frac{1}{n}\lambda I \to 0$ (as $n \to \infty$))

$$\approx \frac{\sigma^2}{n} (C + 0)^{-1}$$

$$= \frac{\sigma^2}{n} C^{-1}$$



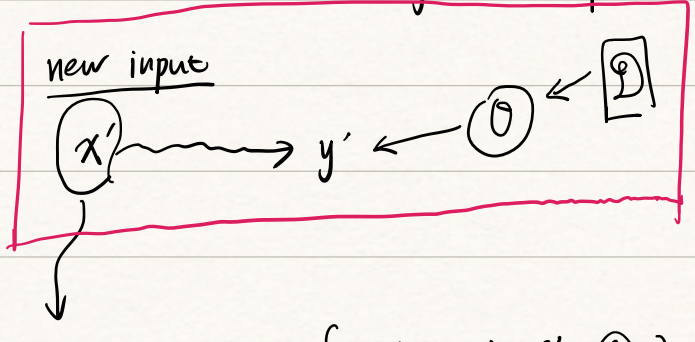
this means when we have MORE samples,

the uncertainty about our θ decays on $O(\frac{1}{n})$

⇓

We have a distribution (not just a point estimate) for $\hat{\theta}_{MAP}$

Predictions:



$$P(y'|x', \mathcal{D}) = \int P(y', \theta | x', \mathcal{D}) \, d\theta$$

$$= \int P(y'|x', \theta) P(\theta) \, d\theta$$

posterior $P(\theta|\mathcal{D})$

$P(y'|x', \mathcal{D})$ ← distribution over target

When x' is the test sample

likelihood prior

How to select a model?

↓

Suppose that there are 2 Competing Models!

$\mathcal{F}_1$: $P(y|x, \theta^1, \sigma^2) = \mathcal{N}(y; \theta^T \phi^{(1)}(x), \sigma^2)$ $\quad \theta \in \mathbb{R}^{d_1}$

$\mathcal{F}_2$: $P(y|x, \theta^2, \sigma^2) = \mathcal{N}(y; \theta^T \phi^{(2)}(x), \sigma^2)$ $\quad \theta \in \mathbb{R}^{d_2}$

$\phi^{(1)} \rightsquigarrow$ a linear model $\quad \phi^{(1)}(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$

$\phi^{(2)} \rightsquigarrow$ a quadratic moder $\quad \phi^{(2)}(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_1^2 \end{bmatrix}$

$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$\mathcal{F}_1 \subseteq \mathcal{F}_2$

If can computer model evidence / Bayesian score $P(\mathcal{D}|\mathcal{F}_i)$

↓ 看到里的可能性

we should select the larger Model Evidence!

$$i^* = \underset{i=1,\dots K}{\operatorname{argmax}} \; P(\mathcal{D}|\mathcal{F}_i)$$

Difficult to Compute

→ BIC

Method: Bayesian Information Criterion $\approx P(\mathcal{D}|\mathcal{F}_i)$

Conclusion: works well when d_i is small and $n \to \infty$

When $n \to \infty$, under some regularity conditions on the likelihood and prior terms

model evidence — log-likelihood + penalty

$$i^* = \arg\max_{i=1,2,\dots K} P(\mathcal{D}|\mathcal{F}_i) \approx BIC(i)$$

$$BIC(i) = \ell(\mathcal{D}; \hat{\theta}_i) - \frac{d_i}{2}\log n$$

$\mathcal{D}$: dataset; $\hat{\theta}_i$: maximize log-likelihood under model $\mathcal{F}_i$

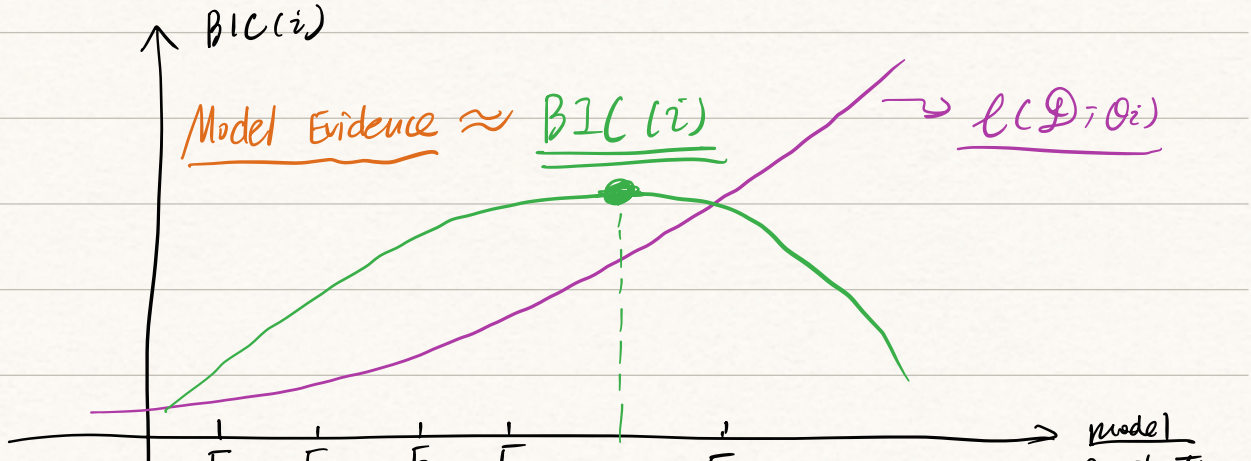
$\frac{d_i}{2}\log n$: penalization terms; d_i: # of free parameters in $\mathcal{F}_i$

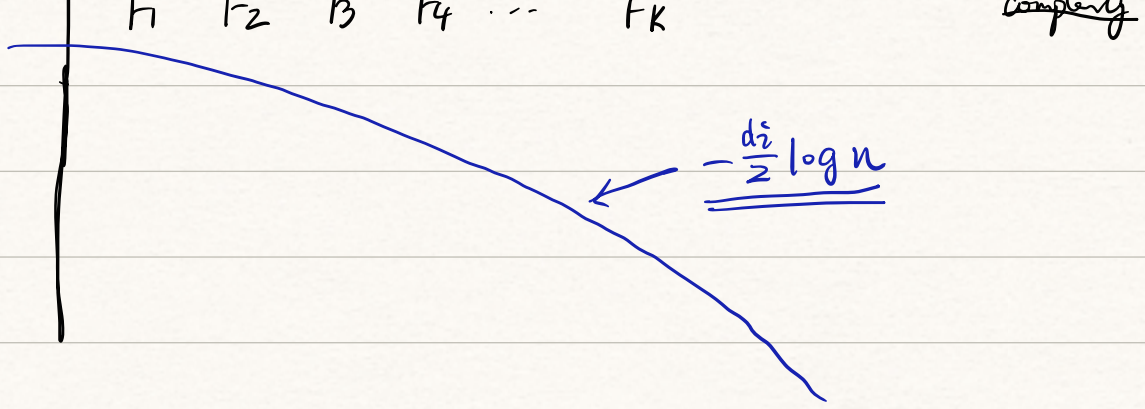
$$\hat{\theta}_i = \arg\max \ell(\mathcal{D}; \theta)$$

$\ell(\mathcal{D}; \theta)$: log-likelihood

Intuition: i) BIC penalizes models with a Large # of parameters; $d_i \uparrow$, $BIC(i) \downarrow$

ii) without the complexity penalty terms ($\frac{d_i}{2}\log n$), we will always choose the most complex model





Example: ① $\ell_1 = -100$ $\ell_2 = -50$

ML

$\ell_i = \ell(\mathcal{D}; \hat{\theta}_i)$

Negative → $\log P < 0$

$\ell < 0$

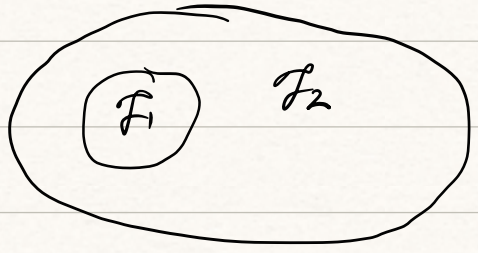
② Linear Model $d_1 = d+1$

Quad. Model $d_2 = \binom{d+2}{2} = \frac{1}{2}(d+1)(d+2)$

1. Why is $\tilde{\ell}_1 \le \tilde{\ell}_2$ for $\mathcal{F}_1 \subseteq \mathcal{F}_2$?

Pf: $\tilde{\ell}_1 = \underset{\theta, \mathcal{F}_1}{\arg\max} \log L(\theta; \mathcal{D})$

$\tilde{\ell}_2 = \underset{\theta, \mathcal{F}_2}{\arg\max} \log L(\theta; \mathcal{D})$



$\Rightarrow \tilde{\ell}_1 \le \tilde{\ell}_2$

2. Prefer Model 2 over Model 1 (when)

i.e. $BIC(2) > BIC(1)$

$$BIC(2) = \hat{\ell}_2 - \frac{d_2}{2}\log n = -50 - \frac{(d+1)(d+2)}{4}\log n$$

$$BIC(1) = \hat{\ell}_1 - \frac{d_1}{2}\log n = -100 - \frac{d+1}{2}\log n$$

- if n small ⇝ choose Model 2.
- if n big ⇝ choose Model 1

#