Lecture 6

Regularization

MLE can overfit the training data.

Ex. $\vec{\theta} = (\theta_h, 1-\theta_h)$

$\uparrow$ heads in coin tosses

$$\mathcal{D} = \{\underbrace{h, h, h}_{N=3}\} \Rightarrow \theta_h^* = \frac{3}{3} = 1$$

Now, $Ber(y|\vec{\theta}^*)$ will predict no tails $\Rightarrow$ the model does not generalize

To prevent overfitting, one uses regularization:

$$\mathcal{J}(\vec{\theta}; \lambda) = + \sum_{n=1}^{N} \log p(\vec{y}_n | \vec{x}_n, \vec{\theta}) + \lambda \underbrace{C(\vec{\theta})}_{\text{complexity penalty}}$$

If $\lambda = 1$ and $C(\vec{\theta}) = \log \underbrace{p(\vec{\theta})}_{\text{prior}}$, we have

$$\mathcal{J}(\vec{\theta}; \lambda) = \sum_{n=1}^{N} \log p(\vec{y}_n | \vec{x}_n, \vec{\theta}) + \log p(\vec{\theta}) =$$

$$= \log p(\mathcal{D}|\vec{\theta}) + \log p(\vec{\theta}) = \log p(\vec{\theta}|\mathcal{D}) + \text{const}$$

MAP estimate: $\vec{\theta}_{MAP} = \underset{\vec{\theta}}{\operatorname{argmax}} \log p(\vec{\theta}|\mathcal{D})$

<u>Ex.</u> Coin tossing:

use $p(\theta) = \mathrm{Beta}(\theta | a, b)$, where

$$\mathrm{Beta}(\theta | a, b) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$ is

the <u>beta distribution</u>.

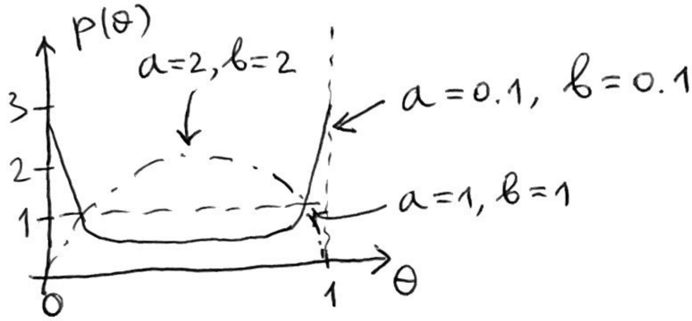
Here, $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the

beta function, and

$$\Gamma(a) = \int^{\infty} dx\, x^{a-1} e^{-x}.$$

gamma function

Note that $a, b > 0$; if $a = b = 1$, the beta distribution is uniform.



$$\underbrace{E[\theta]}_{\text{mean}} = \frac{a}{a+b}$$

We need to maximize

$$\log p(\mathcal{D}|\vec{\theta}) + \log p(\vec{\theta}) \underset{\text{maximize}}{\Rightarrow}$$

$$\Rightarrow \underbrace{N_1}_{\#\text{ heads}} \log\theta + \underbrace{N_0}_{\#\text{ tails}} \log(1-\theta) + (a-1)\log\theta + (b-1)\log(1-\theta).$$

$$N = N_0 + N_1$$

$$\frac{\partial}{\partial\theta}\log p(\theta|\mathcal{D})\Big|_{\theta_{MAP}} = \frac{N_1+(a-1)}{\theta_{MAP}} - \frac{N_0+(b-1)}{1-\theta_{MAP}} = 0,$$ or

$$[N_1+(a-1)](1-\theta_{MAP}) = [N_0+(b-1)]\,\theta_{MAP},$$

$$\theta_{MAP} = \frac{N_1+(a-1)}{N+a+b-2}.$$

we might set $a=b=2 \Rightarrow \theta_{MAP} = \frac{N_1+1}{N+2}$
"add one smoothing"

Ex. Regression:

Use zero-mean gaussian for $p(\vec{w})$:

maximize $\underbrace{\log p(\mathcal{D}|\vec{w})}_{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f(\vec{x}_i;\vec{w},b))^2 + \text{const}(\vec{w})} - \lambda|\vec{w}|^2$ wrt $\vec{w}$

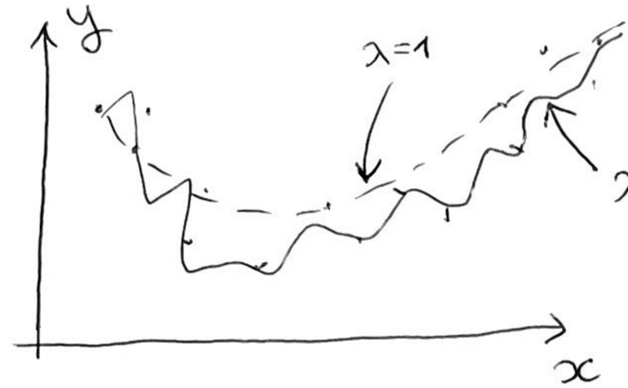
λ: $\frac{1}{2}\times$precision prm; $\log p(\vec{w}) = -\lambda|\vec{w}|^2 + \text{const}(\vec{w})$

$$|\vec{w}|^2 = \sum_{d=1}^{D} w_d^2 \qquad \vec{x}\in R^D$$

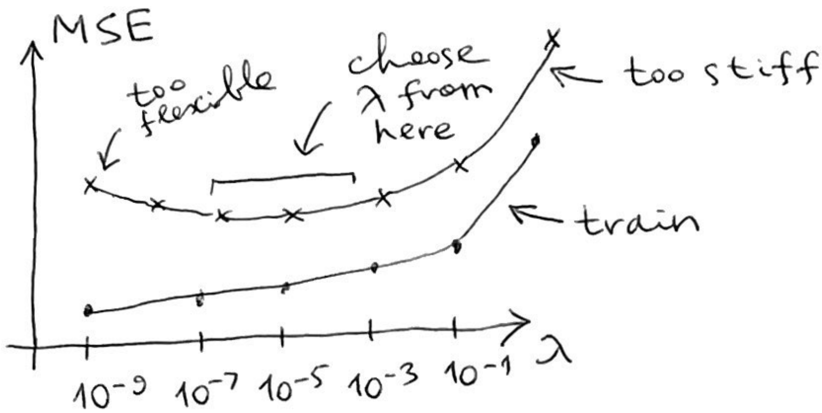
Linear regression:

$$f(\vec{x};\vec{w},b) = \sum_{d=1}^{D} w_d x_d + b.$$

Choice of λ:



Here,
$$f(x;\vec{w},b) = \sum_{d=1}^{14} w_d x^d + b$$



Bayesian statistics

$$\underbrace{p(\vec{\theta}|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{p(\mathcal{D}|\vec{\theta})}^{\text{likelihood}}\,\overbrace{p(\vec{\theta})}^{\text{prior}}}{\underbrace{\int d\vec{\theta}'\,p(\mathcal{D}|\vec{\theta}')\,p(\vec{\theta}')}_{p(\mathcal{D}),\ \text{marginal likelihood}}}$$

Supervised: $\mathcal{D} = \{(\vec{x}_n, \vec{y}_n)\}_{n=1}^{N}$

Unsupervised: $\mathcal{D} = \{\vec{y}_n\}_{n=1}^{N}$

Once we have $p(\vec{\theta}|\mathcal{D})$, we can compute the <u>predictive distribution</u>

$$p(\vec{y}|\vec{x}, \mathcal{D}) = \int d\vec{\theta}\; p(\vec{y}|\vec{x}, \vec{\theta})\, p(\vec{\theta}|\mathcal{D}) .$$

↑ supervised case

For example, if $p(\vec{\theta}|\mathcal{D}) = \delta(\vec{\theta} - \vec{\theta}_{MAP})$,

$$p(\vec{y}|\vec{x}, \mathcal{D}) = \underbrace{p(\vec{y}|\vec{x}, \vec{\theta}_{MAP})}_{\text{a single model predominates}}$$

<u>Conjugate priors</u>: priors that are parameterized similarly to the likelihood.

<u>Ex.</u>
$$\begin{cases} p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0} & \text{likelihood} \\ p(\theta) = \mathrm{Beta}(\theta|a,b) = \dfrac{1}{B(a,b)}\,\theta^{a-1}(1-\theta)^{b-1} & \text{prior} \end{cases}$$

Then $p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|a+N_1, b+N_0)$.

In particular, $$E[\theta] = \frac{a+N_1}{\underbrace{a+b}_{\tilde{N} = \text{equiv. sample size}} + \underbrace{N_1+N_0}_{N = \text{sample size}}}$$

One can write $m = \dfrac{a}{\tilde{N}}$

$$E[\theta] = \frac{\tilde{N}\frac{a}{\tilde{N}} + N\frac{N_1}{N}}{N+\tilde{N}} = \underbrace{\frac{\tilde{N}}{N+\tilde{N}}}_{\lambda} m + \frac{N}{N+\tilde{N}}\underbrace{\frac{N_1}{N}}_{\theta^*,\ \text{MLE}} = \lambda m + (1-\lambda)\,\theta^*$$

as $N \uparrow$, $\lambda \downarrow$ and $\underbrace{E[\theta]}_{\substack{\text{posterior} \\ \text{mean}}} \to \theta^*$.

predictive distribution:

$$p(y=1|\mathcal{D}) = \int_0^1 d\theta \, \underbrace{p(y=1|\theta)}_{\theta} \, p(\theta|\mathcal{D}) = E[\theta|\mathcal{D}] =$$

$$= \frac{N_1 + a}{N + a + b}.$$

If $a = b = 1$ (uniform prior), we have:

$$p(y=1|\mathcal{D}) = \frac{N_1 + 1}{N + 2}$$

pseudocounts