Lecture 9

Bayesian information criterion (BIC)

Recall that under the Laplace approximation,

$$p(\vec{\theta}|\mathcal{D}, M) \simeq \frac{p(\vec{\theta}, \mathcal{D}|M)}{\underbrace{p(\mathcal{D}|M)}_{\text{" } Z \text{ from before}}}, \text{ where}$$

$$p(\mathcal{D}|M) = e^{\log p(\hat{\vec{\theta}}, \mathcal{D}|M)} (2\pi)^{D/2} |H|^{-1/2}.$$

$$D = \dim\{\vec{\theta}\}$$

Using $\hat{\vec{\theta}} = \vec{\theta}_{MAP}$, we obtain:

$$\log p(\mathcal{D}|\overset{\text{model type}}{M}) = \underbrace{\log p(\mathcal{D}|\vec{\theta}_{MAP}, M)}_{\text{log-likelihood}} + \underbrace{\log p(\vec{\theta}_{MAP}|M) - \frac{1}{2}\log|H|}_{\text{penalty terms}} + \underbrace{\text{const}}_{\substack{\text{indep. of } M,\\ \text{indep. of } N}}$$

In the large-N limit, $\vec{\theta}_{MAP} \to \underbrace{\vec{\theta}^*}_{\text{MLE}}$ and the effects of the prior are negligible, such that

$$\log p(\mathcal{D}|M) \simeq \log p(\mathcal{D}|\vec{\theta}^*, M) - \underbrace{\frac{1}{2}\log|H|}_{\text{Occam factor}}, \text{ where}$$

$$H_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j}\left[-\log p(\mathcal{D}|\vec{\theta}, M) - \underbrace{\log p(\vec{\theta}|M)}_{\text{discard the prior}}\right]\Bigg|_{\vec{\theta}^*} \ominus$$

$$\textcircled{=} - \frac{\partial^2}{\partial\theta_i \partial\theta_j} \underbrace{\log p(\mathcal{D}|\vec{\theta}, M)}_{\sum_{k=1}^{N \leftarrow \text{sample size}} \log p(\mathcal{D}_k|\vec{\theta}, M)\text{, additive over data points}} \Big|_{\vec{\theta}^*}.$$

We can write

$$H = \sum_{k=1}^{N} H_k = N \langle H \rangle \text{ , then}$$

$$\log|H| = \log|N\langle H\rangle| \underset{\uparrow \text{ if } H \text{ is full rank}}{=} \log(N^D \langle H \rangle) \textcircled{=}$$

$$\textcircled{=} D \log N + \underbrace{\log\langle H\rangle}_{\text{discard, indep. of } N}.$$

Thus, $\underbrace{\log p(\mathcal{D}|M)}_{\text{BIC score}} \simeq \log p(\mathcal{D}|\vec{\theta}^*, M) - \underbrace{\frac{D_M}{2} \log N}_{\text{model complexity penalty}}.$

D_M = # parameters in model M

<u>The issue</u> <u>of rank</u>: suppose that ($D_M = 3$)

$$\underbrace{H_1}_{\text{due to likelihood only}} = \begin{pmatrix} H_{11} & H_{12} & 0 \\ H_{21} & H_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} N\langle H_{11}\rangle & N\langle H_{12}\rangle & 0 \\ N\langle H_{21}\rangle & N\langle H_{22}\rangle & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and that $-\log p(\vec{\theta}|M) = \lambda|\vec{\theta}|^2$, then

$$H = H_y + \lambda \mathbb{I} = \begin{pmatrix} N\langle H_{11}\rangle \underbrace{+\lambda}_{\text{neglect}} & N\langle H_{12}\rangle & 0 \\ N\langle H_{21}\rangle & N\langle H_{22}\rangle \underbrace{+\lambda}_{\text{neglect}} & 0 \\ 0 & 0 & \lambda \end{pmatrix}.$$

Now, $\det H \approx N^2\langle H_{11}\rangle\langle H_{22}\rangle\lambda - N^2\langle H_{12}\rangle\langle H_{21}\rangle\lambda =$

$= N^2\lambda\left[\langle H_{11}\rangle\langle H_{22}\rangle - \langle H_{12}\rangle\langle H_{21}\rangle\right].$

Thus,

$$\log|H| = \underset{\substack{\uparrow \\ \text{not } 3!}}{2}\log N + \underbrace{\text{const}(N)}_{\substack{\text{discard} \\ \text{as before}}}$$

Information theory

Entropy of a discrete random variable X:

$$H(X) = -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k) =$$
$$= -E[\log p(X)].$$

If $\log_2$ is used, units of entropy are bits.

Max entropy: $p(X=k) = \frac{1}{K}$, $\forall k$.

Then $H(X) = \log_2 K$.

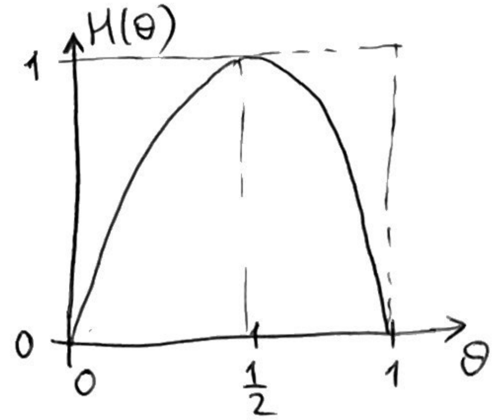
Min entropy: $p(X=k') = 1$, all others $= 0$.

Then $H(X) = 0$.

If $K=2$ (binary data), we have:

$$\begin{cases} p(X=1)=\theta, \\ p(X=0)=1-\theta. \end{cases}$$

Then $H(X) = -\theta \log_2 \theta - (1-\theta)\log_2(1-\theta) = H(\theta)$



$\max\{H(\theta)\} = 1$ bit

Cross-entropy between discrete distributions p & q_θ: $H(p,q_\theta) = -\sum_{k=1}^{K} p_k \log_2 q_{\theta k}$

Joint entropy:

$$H(X,Y) = -\sum_{x,y} p(x,y)\log_2 p(x,y)$$

If X & Y are indep., $p(x,y) = p(x)p(y)$

and $H(X,Y) = H(X) + H(Y)$ upper bound on entropy

If X & Y are correlated, we expect

$$H(X,Y) < H(X) + H(Y).$$

Lower bound: $H(X,Y) \geq \max\{H(X), H(Y)\} \geq 0$. (*)

Conditional entropy:

$$H(Y|X) = -\sum_x p(x) \sum_y p(y|x) \log p(y|x) =$$

$$= -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)} = -\sum_{x,y} p(x,y) \log p(x,y) +$$

$$+ \sum_x p(x) \log p(x) = H(X,Y) - H(X).$$

Thus, $H(X,Y) = H(X) + H(Y|X)$.

Likewise, $H(X,Y) = H(Y) + H(X|Y)$.

Since $H(X|Y) \geq 0$, $H(Y|X) \geq 0$, we recover Eq. $(*)$.

Moreover,

$$\underbrace{H(X,Y)}_{H(X)+H(Y|X)} \leq H(X) + H(Y) \Rightarrow H(Y|X) \leq H(Y)$$

If X completely determines Y: $H(Y|X) = 0$

If X & Y are indep.: $H(Y|X) = H(Y)$.

"Conditioning on data does not increase uncertainty, <u>on</u> <u>average</u>".

→ Chain rule for entropy:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \ldots, X_{i-1}) =$$

$$= H(X_1) + H(X_2|X_1) + \ldots + H(X_n | X_1, \ldots, X_{n-1})$$

Continuous random variables:

Differential entropy:

$$h(X) = -\int dx\, p(x) \log p(x).$$

For example, if $X \sim \mathcal{U}(0,a)$,

$$h(X) = -\int_0^a dx\, \frac{1}{a} \log \frac{1}{a} = \log a,$$

can be <0 if $a<1$.

Similarly, if $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$h(X) = -\int dx\, p(x) \left[-(x-\mu)^2/2\sigma^2 - \frac{1}{2}\log(2\pi\sigma^2) \right] =$$

$$= \frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2).$$

KL divergence:

$$D_{KL}(p\|q) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

or

$$D_{KL}(p\|q) = \int dx\, p(x) \log \frac{p(x)}{q(x)}.$$

Note that $D_{KL}(p\|q) = \underbrace{-\sum_k p_k \log q_k}_{H(p,q),\ \text{cross entropy}} +$

$$+ \underbrace{\sum_k p_k \log p_k}_{-H(p),\ \text{entropy of } p}$$

Ex. $$D_{KL}(p,q) = -\underbrace{H(p)}_{\frac{\log(2\pi\sigma_1^2)+1}{2}} - \int dx\, p(x)\left[-\frac{(x-\mu)^2}{2\sigma_2^2} - \frac{1}{2}\log(2\pi\sigma_2^2)\right] \quad (11)$$

$$\begin{cases} p = \mathcal{N}(\mu, \sigma_1^2), \\ q = \mathcal{N}(\mu, \sigma_2^2). \end{cases}$$

$$= -\frac{1}{2} - \frac{1}{2}\log(2\pi\sigma_1^2) + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{1}{2}\log(2\pi\sigma_2^2) =$$

$$= \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2}.$$

If $\sigma_2 \gg \sigma_1$, $D_{KL} \simeq \log\frac{\sigma_2}{\sigma_1} - \frac{1}{2} \gtrsim 1$.

If $\sigma_2 \ll \sigma_1$, $D_{KL} \simeq \frac{\sigma_1^2}{2\sigma_2^2} \gg 1$.

In fact, $D_{KL} \geq 0$, with $D_{KL} = 0$ iff $p = q$.

Recall Jensen's inequality:

for any convex function f,

$$f\left(\sum_{i=1}^{n} \lambda_i \vec{x}_i\right) \leq \sum_{i=1}^{n} \lambda_i f(\vec{x}_i)$$

$$0 \leq \lambda_i \leq 1, \quad \sum_{i=1}^{n} \lambda_i = 1.$$

"f of the average is less than the average of f's"

Now, consider $$-D_{KL}(p\|q) = \sum_x p(x)\log\frac{q(x)}{p(x)} \leq$$

$$\leq \log\left(\sum_x p(x)\frac{q(x)}{p(x)}\right) = \log 1 = 0.$$

Jensen; log is concave

$\hookrightarrow$ $D_{KL}(p\|q) \geq 0$.

→ Uniform distribution maximizes the entropy:

$$u(x) = \frac{1}{K}$$ ← # outcomes

Then $$0 \leq D_{KL}(p \,\|\, u) = \sum_x p(x) \log \frac{p(x)}{u(x)} =$$

$$= \log K - H(p), \text{ or}$$

$$H(p) \leq \log K.$$