Notes on Learning from Data, MLE and MAP

① Chain Rule

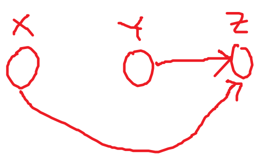
$$p(x_1, \ldots, x_D) = p(x_1) \prod_{i=2}^{D} p(x_i \mid x_1, \ldots, x_{i-1})$$

$$
\begin{aligned}
p(x, y, z) &= p(x \mid y, z)\, p(y \mid z)\, p(z) \quad ① \\
&= p(x \mid y, z)\, p(z \mid y)\, p(y) \quad ② \\
&= p(y \mid x, z)\, p(x \mid z)\, p(z) \quad ③ \\
&= p(y \mid x, z)\, p(z \mid x)\, p(x) \quad ④ \\
&= p(z \mid x, y)\, p(x \mid y)\, p(y) \quad ⑤ \\
&= p(z \mid x, y)\, p(y \mid x)\, p(x) \quad ⑥
\end{aligned}
$$

In general, suppose that we have N Random variables $X_1, \ldots, X_N$ the total number T of possible joint distribution $p(X_1, \ldots, X_N)$ factorizations using chain-rule can be written as:

$$T = N \cdot (N-1)(N-2) \cdots 2 \cdot 1$$
$$= N!$$

② Marginal Independence



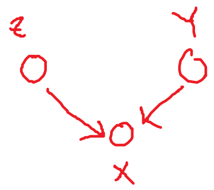
Choose the factorization of $p(x, y, z)$ that fits the graphical model

$$
\begin{aligned}
⑥\; p(x, y, z) &= p(z \mid x, y)\, p(y \mid x)\, p(x) \\
&= p(z \mid y, x)\, p(y)\, p(x)
\end{aligned}
$$

$$⑤\; p(x, y, z) = p(z \mid x, y)\, p(x \mid y)\, p(y)$$

$$= p(z \mid x, y)\, p(x)\, p(y)$$

③ Conditional Independence



To prove $Z \perp Y \mid X$,
Just show that either:

① $p(z \mid x, y) = p(z \mid x)$
$p(y \mid x, z) = p(y \mid x)$

② $p(y, z \mid x) = p(y \mid x)\, p(z \mid x)$

Ex. Show that $Z \not\perp Y \mid X$

$$p(z \mid x, y) = \frac{p(x, y, z)}{p(x, y)} = \frac{p(x \mid y, z) \cdot \cancel{p(y)}\, p(z)}{p(x \mid y)\, \cancel{p(y)}}$$

$$\neq p(z \mid x)$$

Further Reading:

1. Chapter 10 Probabilistic Machine Learning, Murphy 2012
2. Chapter 20 Information Theory, Inference, and Learning Algorithm, Mackay 2003

④ Mean and Variance

Mean $\Rightarrow \mathbb{E}_{x \sim p}[X]$. Suppose that $x \sim \text{Bernoulli}(\theta)$, that is $p(x) = \theta^x (1-\theta)^{(1-x)}$ with $x \in \{0, 1\}$. Then:

$$\mathbb{E}_{x \sim p}[x] = \sum_{x=0}^{1} p(x) \cdot x$$

$$= 0 \cdot p(x=0) + 1 \cdot p(x=1)$$

$$= \theta$$

Variance $\Rightarrow$

$$\mathbb{E}_{x \sim p}\left[\left(X - \mathbb{E}_{x \sim p}[x]\right)^2\right] = \sum_{x=0}^{1} (x-\theta)^2 \cdot p(x)$$

$$= (0-\theta)^2 \cdot p(x=0) + (1-\theta)^2 \cdot p(x=1)$$

$$= \theta^2 (1-\theta) + (1-\theta)^2 \cdot \theta$$

$$= \theta(1-\theta)(\theta + 1 - \theta)$$

$$= \theta(1-\theta)$$

5 Entropy and KL-divergence

Entropy: $H(X) = \mathbb{E}_{x\sim p}[I(x)] = \mathbb{E}_{x\sim p}[-\log p(x)]$

Ex: $x \sim \text{Bernoulli}(\theta) \Rightarrow H(x) = -\sum_{x=0}^{1} p(x) \cdot \log(p(x))$

$$= -[\ p(x=0)\ \log p(x=0) + p(x=1) \cdot \log p(x=1)]$$

$$= -[(1-\theta)\cdot\log(1-\theta) + \theta \cdot \log\theta]$$

$$= -[\log(1-\theta)^{(1-\theta)} + \log\theta^{\theta}]$$

$$= -[\log \cdot \theta^{\theta} \cdot (1-\theta)^{(1-\theta)}]$$

KL-divergence: $KL[p \,\|\, Q] = \mathbb{E}_{x\sim p(x)}\left[\log \frac{P(x)}{Q(x)}\right]$

Properties of KL-divergence:

1 $KL[p \,\|\, Q] \geq 0$ (Gibb's inequality)

Proof:

Jensen's inequality states that if f is a convex function and X is a random variable, then f satisfies:

$$\mathbb{E}_{x\sim p}[f(x)] \geq f(\mathbb{E}_{x\sim p}[X])$$

Since log is a concave function, then $-\log$ is a convex function.

Setting $f(x) = -\log \frac{Q(x)}{p(x)}$ gives us:

$$KL[p(x) \,\|\, Q(x)] = \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{Q(x)}\right]$$

$$= \mathbb{E}_{x\sim p}\left[-\log \frac{Q(x)}{p(x)}\right]$$

$$\geq -\log \mathbb{E}_{x\sim p}\left[\frac{Q(x)}{p(x)}\right]$$

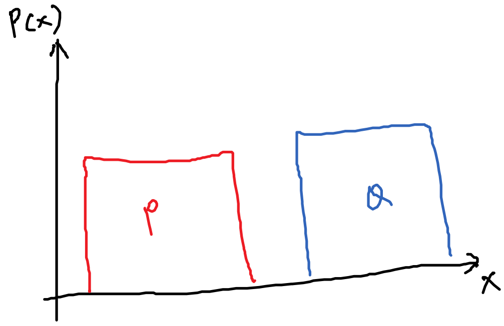
$$= -\log \int_{\mathcal{X}} p(x) \cdot \frac{Q(x)}{p(x)}\, dx$$

$$= -\log \int_{\mathcal{X}} Q(x)\, dx = -\log 1 = 0$$

$p = Q$ iff $KL[p \,\|\, Q] = 0$

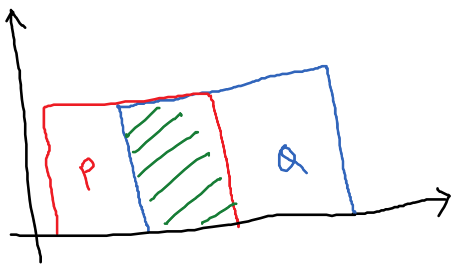
② $KL[P \| Q] \neq KL[Q \| P]$

$$\int p(x) \cdot \log \frac{p(x)}{Q(x)} dx \neq -\int Q(x) \log \frac{p(x)}{Q(x)} dx$$

① 

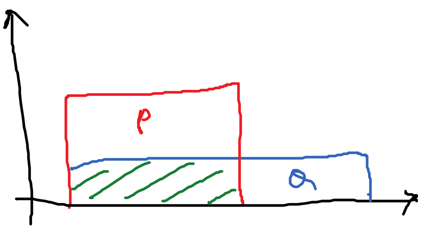
Support (p) and support (Q) is disjoint

Therefore $KL[P \| Q] = KL[Q \| P] = \infty$

② 

Support (p) and support (Q) overlap, but neither is a subset of another. Therefore

$$KL[p \| Q] = KL[Q \| P] = \infty$$

③ 

support (Q) $\subseteq$ Support (p). Therefore

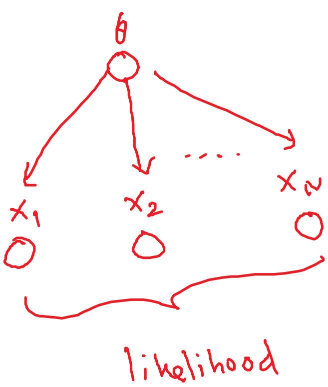
$KL[P \| Q]$ is finite

$KL[Q \| P] = \infty$

Further reading:

- Bishop 2006, chapter 1: Information theory

⑥ MLE



- $x_i \perp x_j \mid \theta \quad \forall i,j : i \neq j \quad 1 \leq i,j \leq N$
- $\mathcal{D} = \{x_i\}_{i=1}^{N}$
- $\hat{\theta}_{MLE} = \max_\theta \log p(\mathcal{D} \mid \theta)$

$$= \max_\theta \log \prod_{i=1}^{N} p(x_i \mid \theta)$$

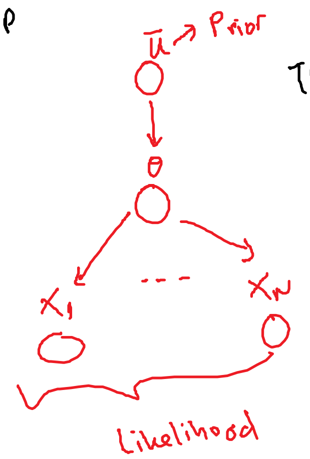
$$= \max_\theta \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

likelihood

$$= \max_{\theta} \sum_{i=1}^{N} \log p(x_i|\theta)$$

$$\approx \frac{\partial \left[\sum_{i=1}^{N} \log p(x_i|\theta) \right]}{\partial \theta} \Bigg|_{\theta} = 0$$

⑦ MAP



The idea comes from Bayesian inference

① Compute posterior

$$P(\theta | \mathcal{D}) = \frac{p(\mathcal{D}|\theta) \cdot p(\theta)}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\theta) \cdot P(\theta)}{\int p(\mathcal{D}|\theta) \cdot p(\theta) \, d\theta}$$

② Given new data x^*, compute predictive posterior

$$P(x^* | \mathcal{D}) = \int p(x^* | \theta) \cdot p(\theta | \mathcal{D}) \, d\theta$$

Drawback of Bayesian → intractability → why? → $P(\mathcal{D})$

Semi-Bayes → discard $p(\mathcal{D})$ → MAP

$$P(\theta | \mathcal{D}) \propto P(\mathcal{D}|\theta) \, p(\theta)$$

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

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

$$\lim_{n \to \infty} \hat{\theta}_{MAP} = \hat{\theta}_{MLE}$$ (Bernstein-Von mises theorem)

Simple observation; suppose that prior is uniform : $p(\theta) = \text{constant}$. Therefore :

$$\hat{\theta}_{MAP} = \max_{\theta} \log p(\mathcal{D}|\theta) + \log p(\theta)$$

$$= \max_{\theta} \log p(\mathcal{D}|\theta) + \text{constant}$$

$$\approx \max_{\theta} \log p(D \mid \theta)$$

$$= \hat{\theta}_{MLE}$$

Bonus: conjugacy

$p(\theta) = U[0,1]$

$P(x \mid \theta) = \text{Bin}(x \mid n, \theta) = \binom{n}{x} \theta^{x} (1-\theta)^{(n-x)}$

$$p(\theta \mid x) \propto p(x \mid \theta) \cdot p(\theta)$$
$$\propto \theta^{x} (1-\theta)^{n-x}$$
$$= \text{Beta}(x+1, n-x+1)$$