香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 8

Naive Bayes, MLE, MAP

Junxian He
Mar 10, 2026

Recap: Generative Models

Recap: Generative Models

X

$p(y \mid x)$

Discriminative

Cat Y

Recap: Generative Models

X

$p(y \mid x)$

Discriminative

Cat Y

Cat Y $p(y)$

Generative $p(x \mid y)$

X

Recap: Generative Models

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

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

If our goal is to predict y, the distribution is often written as:

$$p(y \mid x) \propto p(x \mid y)p(y)$$

$$\begin{aligned} \arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y) \end{aligned}$$

Recap: Naive Bayes

Binary classification: $y \in \{0,1\}, x$ is discrete

Recap: Naive Bayes

Binary classification: $y \in \{0,1\}, x$ is discrete

Consider an email spam detection task, to predict whether the email is spam or not

Recap: Naive Bayes

Binary classification: $y \in \{0,1\}, x$ is discrete

Consider an email spam detection task, to predict whether the email is spam or not

How to represent the text?

Recap: Naive Bayes

Binary classification: $y \in \{0,1\}, x$ is discrete

Consider an email spam detection task, to predict whether the email is spam or not

How to represent the text?

if an email contains the j-th word of the dictionary, then we will set $x_j = 1$; otherwise, we let $x_j = 0$

Recap: Naive Bayes

Binary classification: $y \in \{0,1\}, x$ is discrete

Consider an email spam detection task, to predict whether the email is spam or not

How to represent the text?

if an email contains the j-th word of the dictionary, then we will set $x_j = 1$; otherwise, we let $x_j = 0$

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

Recap: Naive Bayes

Binary classification: $y \in \{0,1\}, x$ is discrete

Consider an email spam detection task, to predict whether the email is spam or not

How to represent the text?

if an email contains the j-th word of the dictionary, then we will set $x_j = 1$; otherwise, we let $x_j = 0$

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

Dimension is the size of the dictionary

Recap: Naive Bayes

Binary classification: $y \in \{0,1\}, x$ is discrete

Consider an email spam detection task, to predict whether the email is spam or not

How to represent the text?

if an email contains the j-th word of the dictionary, then we will set $x_j = 1$; otherwise, we let $x_j = 0$

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

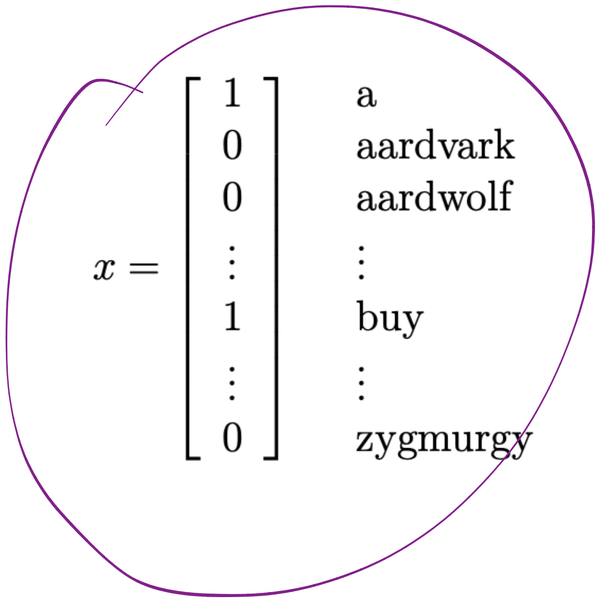
vocabulary

Dimension is the size of the dictionary

Email Spam Classification

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

Email Spam Classification

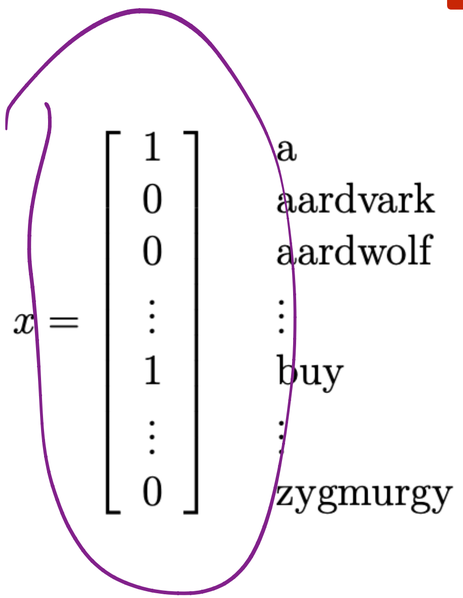


y spam or not x

Suppose the dictionary has 50000 words, how many possible x?

$P(x|y)$ discrete

Email Spam Classification



Suppose the dictionary has 50000 words, how many possible x?

2^{50000}

$P(x|y)$

$x =$ ---

Naive Bayes assumption: x_i's are conditionally independent given y

Email Spam Classification

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

Suppose the dictionary has 50000 words, how many possible x?

Naive Bayes assumption: x_i's are conditionally independent given y

For any i and j, $p(x_i \mid y) = p(x_i \mid y, x_j)$

Email Spam Classification

Email Spam Classification

$$
\begin{aligned}
& p(x_1, \ldots, x_{50000}|y) \\
= \; & p(x_1|y)p(x_2|y, x_1)p(x_3|y, x_1, x_2) \cdots p(x_{50000}|y, x_1, \ldots, x_{49999}) \\
= \; & p(x_1|y)p(x_2|y)p(x_3|y) \cdots p(x_{50000}|y) \\
= \; & \prod_{j=1}^{d} p(x_j|y)
\end{aligned}
$$

chain rule

autoregressive

conditionally independent

Naive

Email Spam Classification

$$
\begin{aligned}
& p(x_1, \ldots, x_{50000}|y) \qquad \text{Autoregressive} \\
&= p(x_1|y)p(x_2|y,x_1)p(x_3|y,x_1,x_2)\cdots p(x_{50000}|y,x_1,\ldots,x_{49999}) \\
&= p(x_1|y)p(x_2|y)p(x_3|y)\cdots p(x_{50000}|y) \\
&= \prod_{j=1}^{d} p(x_j|y)
\end{aligned}
$$

Email Spam Classification

$$
\begin{aligned}
& p(x_1, \ldots, x_{50000} | y) \qquad \text{Autoregressive} \\
&= p(x_1|y)p(x_2|y,x_1)p(x_3|y,x_1,x_2)\cdots p(x_{50000}|y,x_1,\ldots,x_{49999}) \\
&= p(x_1|y)p(x_2|y)p(x_3|y)\cdots p(x_{50000}|y) \\
&= \prod_{j=1}^{d} p(x_j|y)
\end{aligned}
$$

Parameters

$$\phi_{j|y=1} = p(x_j = 1 \,|\, y = 1), \quad \phi_{j|y=1} = p(x_j = 1 \,|\, y = 0), \quad \phi_y = p(y = 1)$$

Email Spam Classification

$$
\begin{aligned}
& p(x_1, \ldots, x_{50000} | y) \qquad \text{Autoregressive} \\
&= p(x_1|y)p(x_2|y,x_1)p(x_3|y,x_1,x_2)\cdots p(x_{50000}|y,x_1,\ldots,x_{49999}) \\
&= p(x_1|y)p(x_2|y)p(x_3|y)\cdots p(x_{50000}|y) \\
&= \prod_{j=1}^{d} p(x_j|y)
\end{aligned}
$$

Parameters

$$\phi_{j|y=1} = p(x_j = 1 \,|\, y = 1), \quad \phi_{j|y=1} = p(x_j = 1 \,|\, y = 0), \quad \phi_y = p(y = 1)$$

50000 x 2 + 1 parameters (dict size is 50000) $<< 2^{50000}$

Maximum Likelihood Estimation

Maximum Likelihood Estimation

$$\mathcal{L}(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) = \prod_{i=1}^{n} p(x^{(i)}, y^{(i)})$$

Maximum Likelihood Estimation

$$\mathcal{L}(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) = \prod_{i=1}^{n} p(x^{(i)}, y^{(i)})$$

$$\begin{aligned}
\phi_{j|y=1} &= \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}} \\
\phi_{j|y=0} &= \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}} \\
\phi_y &= \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}{n}
\end{aligned}$$

Maximum Likelihood Estimation

$$\mathcal{L}(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) = \prod_{i=1}^{n} p(x^{(i)}, y^{(i)})$$

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$

$$\phi_y = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}{n}$$

Count the occurrence of x_j in spam/ non-spam emails and normalize

Prediction

Prediction

$$
\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x)} \\
&= \frac{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1) + \left(\prod_{j=1}^{d} p(x_j|y=0)\right) p(y=0)}
\end{aligned}
$$

Prediction

$$
\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x)} \\
&= \frac{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1) + \left(\prod_{j=1}^{d} p(x_j|y=0)\right) p(y=0)}
\end{aligned}
$$

Naive Classifier

P(y=1|x)

P(y=0|x)

Laplace Smoothing

Laplace Smoothing

What if we never see the word “learning” in training data but “learning” exists in the test data?

Laplace Smoothing

What if we never see the word “learning” in training data but “learning” exists in the test data?

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$

x_j = "learning"

Laplace Smoothing

What if we never see the word “learning” in training data but “learning” exists in the test data?

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$

Suppose the index in the dictionary for “learning” is q

$$p(x_q = 1 \mid y = 1) = 0$$

$$p(x_q = 1 \mid y = 0) = 0$$

Laplace Smoothing

What if we never see the word “learning” in training data but “learning” exists in the test data?

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$

Suppose the index in the dictionary for “learning” is q

$$p(x_q = 1 \,|\, y = 1) = 0$$

$$p(x_q = 1 \,|\, y = 0) = 0$$

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$$

$$= \frac{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1) + \left(\prod_{j=1}^{d} p(x_j|y=0)\right) p(y=0)}$$

Laplace Smoothing

What if we never see the word “learning” in training data but “learning” exists in the test data?

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$

Suppose the index in the dictionary for “learning” is q

$$p(x_q = 1 \,|\, y = 1) = 0$$

$$p(x_q = 1 \,|\, y = 0) = 0$$

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$$

$$= \frac{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1) + \left(\prod_{j=1}^{d} p(x_j|y=0)\right) p(y=0)} = \frac{0}{0}$$

Laplace Smoothing

Laplace Smoothing

Take the problem of estimating the mean of a multinomial random variable z taking values in {1, …, k}. Given the independent observations $\{z^{(1)}, \cdots, z^{(n)}\}$

Laplace Smoothing

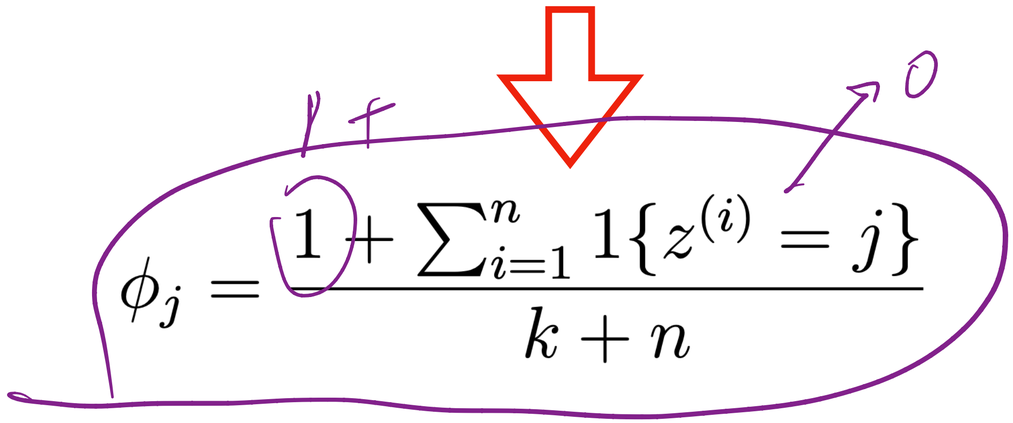
Take the problem of estimating the mean of a multinomial random variable z taking values in {1, …, k}. Given the independent observations $\{z^{(1)}, \cdots, z^{(n)}\}$

$$\phi_j = p(z = j) \qquad \phi_j = \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}{n}$$

Laplace Smoothing

Take the problem of estimating the mean of a multinomial random variable z taking values in {1, …, k}. Given the independent observations $\{z^{(1)}, \cdots, z^{(n)}\}$

$$\phi_j = p(z = j) \qquad \phi_j = \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}{n}$$



Laplace Smoothing

Take the problem of estimating the mean of a multinomial random variable z taking values in {1, …, k}. Given the independent observations $\{z^{(1)}, \cdots, z^{(n)}\}$

$$\phi_j = p(z = j) \qquad \phi_j = \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}{n}$$

$$\phi_j = \frac{1 + \sum_{i=1}^{n} 1\{z^{(i)} = j\}}{k + n}$$

Why adding k to the denominator?

Laplace Smoothing

Take the problem of estimating the mean of a multinomial random variable z taking values in {1, …, k}. Given the independent observations $\{z^{(1)}, \cdots, z^{(n)}\}$

$$\phi_j = p(z = j) \qquad \phi_j = \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}{n}$$

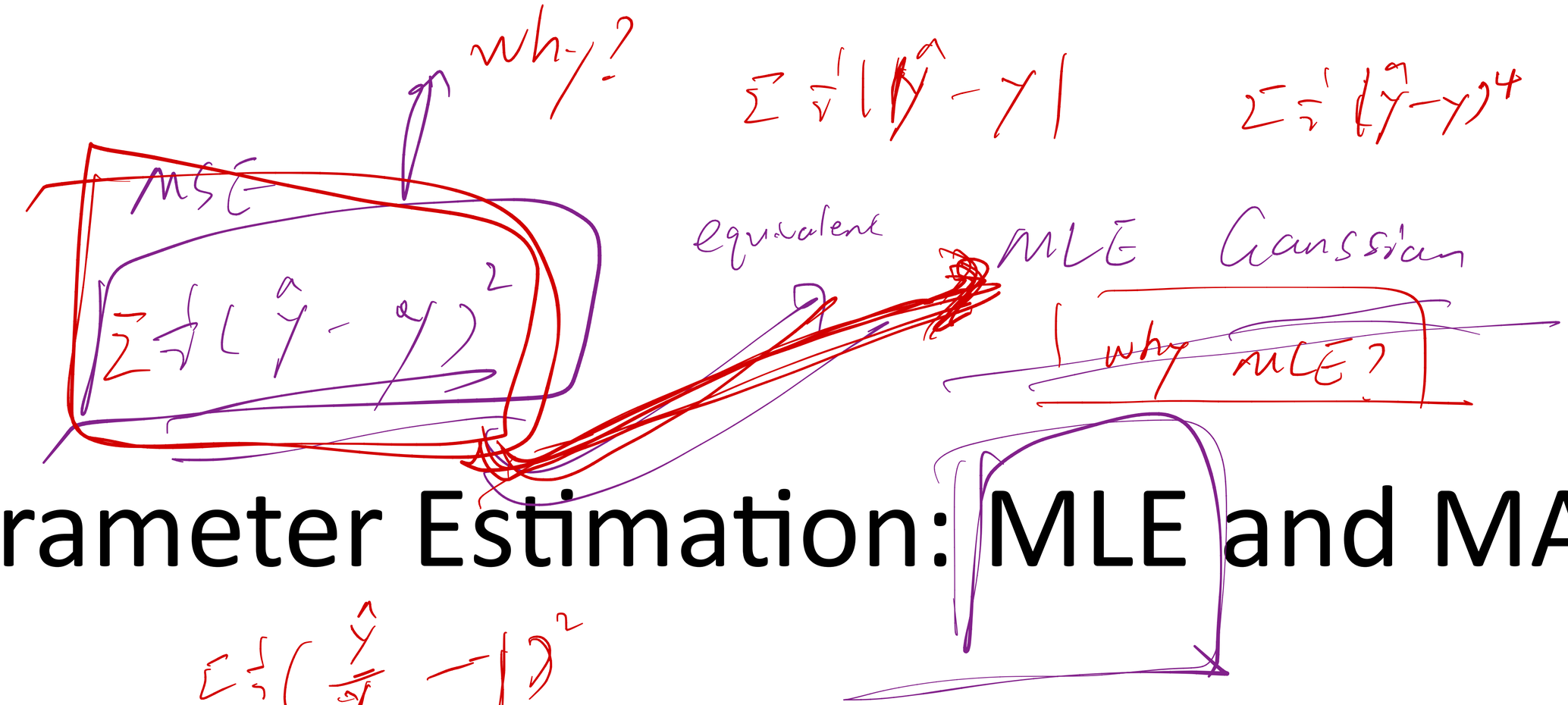
$$\phi_j = \frac{1 + \sum_{i=1}^{n} 1\{z^{(i)} = j\}}{k + n}$$

Why adding k to the denominator?

In the email spam classification case:

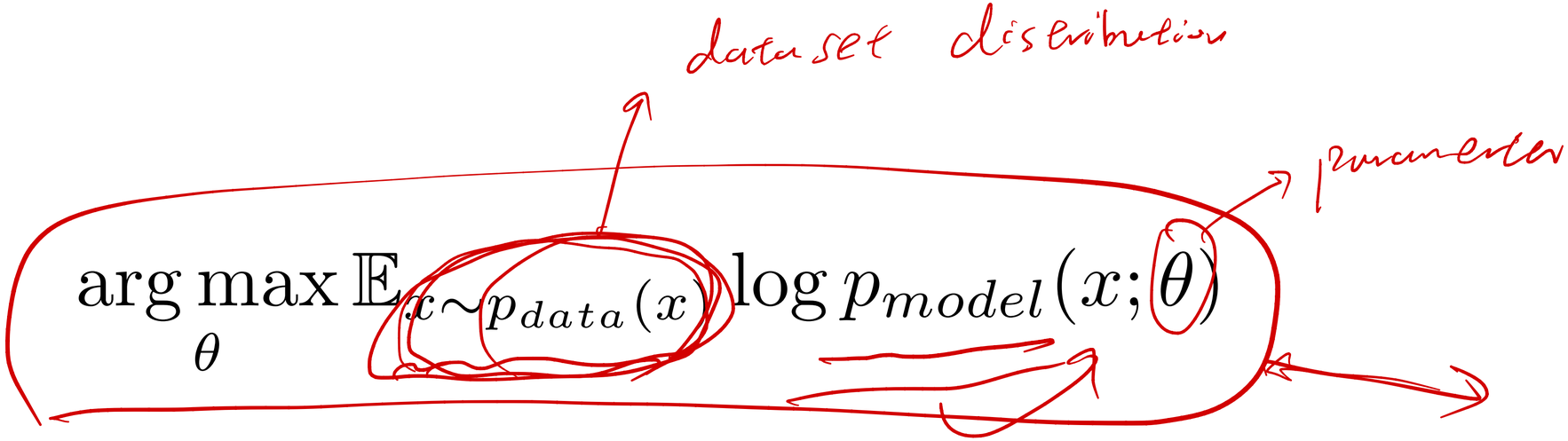
$$\phi_{j|y=1} = \frac{1 + \sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{2 + \sum_{i=1}^{n} 1\{y^{(i)} = 1\}}$$

$$\phi_{j|y=0} = \frac{1 + \sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{2 + \sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$



Parameter Estimation: MLE and MAP

Maximum Likelihood Estimation (MLE)



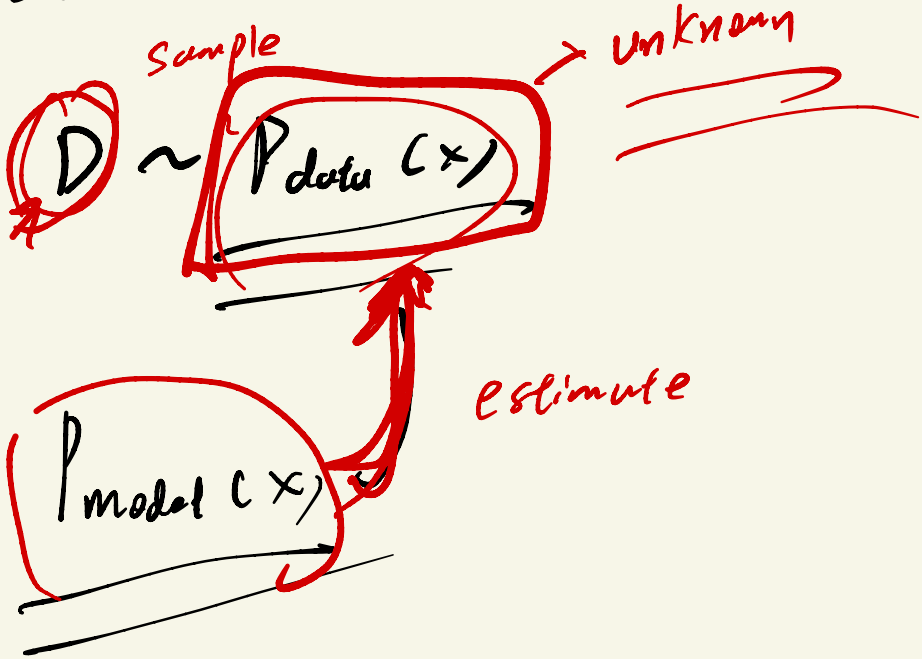
$$\arg\max_{\theta} \mathbb{E}_{x \sim p_{data}(x)} \log p_{model}(x;\theta)$$

In practice:

$$\arg\max_{\theta} \frac{1}{n} \sum_{i}^{n} \log p_{model}(x^{(i)};\theta)$$

$x^{(1)}, x^{(2)}, x^{(n)} \sim p_{data}(x)$

D = dataset



Maximum Likelihood Estimation (MLE)

Suppose $p_{data}(x)$ is the real data distribution, $p_{model}(x;\theta)$ is our model parameterized by θ

$$\arg\max_{\theta} \mathbb{E}_{x\sim p_{data}(x)} \log p_{model}(x;\theta)$$

In practice:

$$\arg\max_{\theta} \frac{1}{n} \sum_{i}^{n} \log p_{model}(x^{(i)};\theta)$$

Maximum Likelihood Estimation (MLE)

Suppose $p_{data}(x)$ is the real data distribution, $p_{model}(x;\theta)$ is our model parameterized by θ

$$\arg\max_{\theta} \mathbb{E}_{x\sim p_{data}(x)} \log p_{model}(x;\theta)$$

In practice:

$$\arg\max_{\theta} \frac{1}{n}\sum_{i}^{n} \log p_{model}(x^{(i)};\theta)$$

$x^{(i)}$ are i.i.d. (independent and identically distributed) samples from $p_{data}(x)$

Maximum Likelihood Estimation (MLE)

Suppose $p_{data}(x)$ is the real data distribution, $p_{model}(x;\theta)$ is our model parameterized by θ

$$\arg\max_{\theta} \mathbb{E}_{x\sim p_{data}(x)} \log p_{model}(x;\theta)$$

In practice:

$$\arg\max_{\theta} \frac{1}{n} \sum_{i}^{n} \log p_{model}(x^{(i)};\theta)$$

Monte Carlo Estimation of Expectation

$x^{(i)}$ are i.i.d. (independent and identically distributed) samples from $p_{data}(x)$

Maximum Likelihood Estimation (MLE)

Suppose $p_{data}(x)$ is the real data distribution, $p_{model}(x;\theta)$ is our model parameterized by θ

$$\arg\max_{\theta} \mathbb{E}_{x \sim p_{data}(x)} \log p_{model}(x;\theta)$$

In practice:

$$\arg\max_{\theta} \frac{1}{n} \sum_{i}^{n} \log p_{model}(x^{(i)};\theta)$$

Monte Carlo Estimation of Expectation

$x^{(i)}$ are i.i.d. (independent and identically distributed) samples from $p_{data}(x)$

Why can we make this approximation?

unbiased estimator:

estimator: $\hat{\theta}$ $\quad\quad$ θ^*

$$E(\hat{\theta}) = \theta^*$$

unbiased

$n \to \infty$

Monte Carlo Estimation of Expectation

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\right] = \mathbb{E}_{x\sim p(x)} f(x)$$

→ real unbiased

$$Var\left[\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\right] = \frac{Var(f(x))}{n}$$

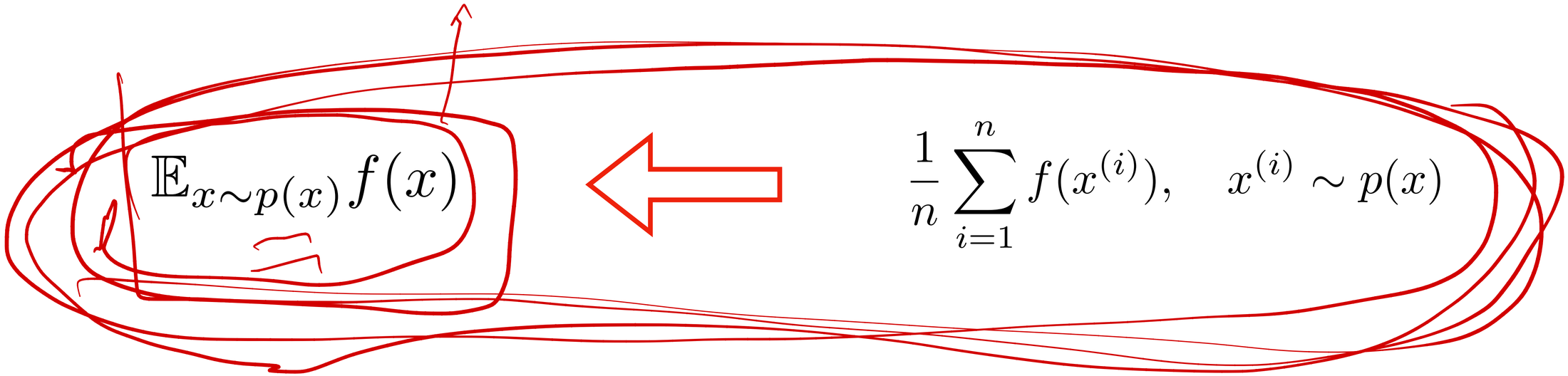
Monte Carlo Estimation of Expectation

$$\mathbb{E}_{x \sim p(x)} f(x)$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\right] = \mathbb{E}_{x \sim p(x)} f(x)$$

$$Var\left[\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\right] = \frac{Var(f(x))}{n}$$

Monte Carlo Estimation of Expectation



$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\right] = \mathbb{E}_{x\sim p(x)} f(x)$$

$$Var\left[\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\right] = \frac{Var(f(x))}{n}$$

Monte Carlo Estimation of Expectation

$$\mathbb{E}_{x\sim p(x)} f(x)$$

$$\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}), \quad x^{(i)} \sim p(x)$$

In practice, n is often small, like 1 sample

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\right] = \mathbb{E}_{x\sim p(x)} f(x)$$

$$Var\left[\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\right] = \frac{Var(f(x))}{n}$$

Sampling and Evaluation of Distributions

Sampling and Evaluation of Distributions

- Some distributions are easy to sample from but hard to compute the probability value (hard to evaluate)
  - Monte Carlo estimation requires this kind of distribution

$P(x)$

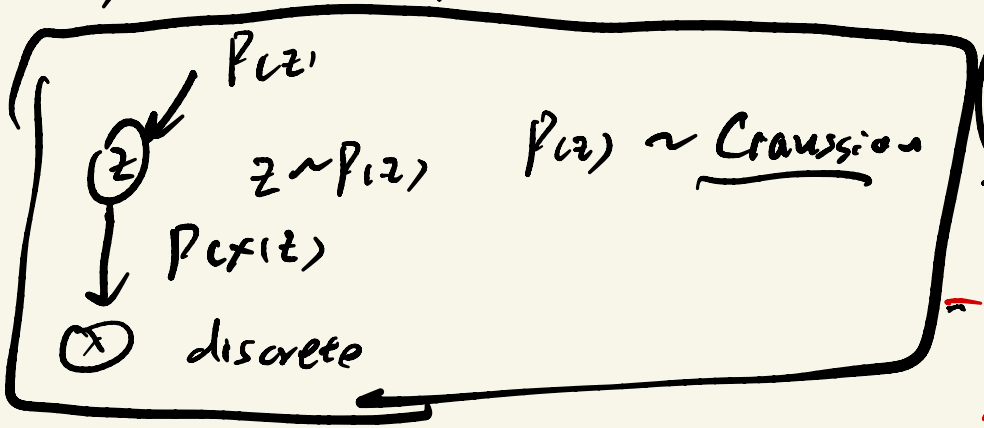
hard to sample but easy to compute

x continuous

$p(x) =$ 

x_0 $\int_z p(x|z)\,p(z)\,dz$

easy to sample but hard to compute



$p(x)$ $x \sim p(x)$

$z_0 \sim p(z)$, $x_0 \sim p(x|z_0)$

f_1 $\int_z p(x_1, z)\,dz$

Sampling and Evaluation of Distributions

- Some distributions are easy to sample from but hard to compute the probability value (hard to evaluate)
  - Monte Carlo estimation requires this kind of distribution

- Some distributions are easy to compute the probability value (easy to evaluate) but hard to sample from
  - How to sample from a distribution efficiently is a separate topic

MLE is Approximating the Real Distribution

$$\arg\max_{\theta} \mathbb{E}_{x \sim p_{data}(x)} \log p_{model}(x; \theta)$$

MLE is Approximating the Real Distribution

$$\arg\max_{\theta} \mathbb{E}_{x \sim p_{data}(x)} \log p_{model}(x; \theta)$$

What is the optimal p_{model} ?

$p_{model} \longrightarrow p_{data}(x)$

MLE is Approximating the Real Distribution

$$\arg\max_{\theta} \mathbb{E}_{x \sim p_{data}(x)} \log p_{model}(x;\theta)$$

What is the optimal p_{model} ?

MLE is equivalent to

$$\arg\min_{\theta} D_{\mathrm{KL}}(p_{data}(x)||p_{model}(x;\theta))$$

MLE is Approximating the Real Distribution

$$\arg\max_{\theta} \mathbb{E}_{x \sim p_{data}(x)} \log p_{model}(x;\theta)$$

What is the optimal p_{model} ?

MLE is equivalent to

KL divergence

$$\arg\min_{\theta} D_{\mathrm{KL}}(p_{data}(x)||p_{model}(x;\theta))$$

$D_{KL} \geq 0$ is a distance metric between two distributions, it is 0 when the two distributions are identical

MLE is Approximating the Real Distribution

$$\arg\max_{\theta} \mathbb{E}_{x \sim p_{data}(x)} \log p_{model}(x;\theta)$$

What is the optimal p_{model} ?

MLE is equivalent to

$$\arg\min_{\theta} D_{\mathrm{KL}}(p_{data}(x)||p_{model}(x;\theta))$$

$D_{KL} \geq 0$ is a distance metric between two distributions, it is 0 when the two distributions are identical

$$D_{\mathrm{KL}}(p(x)||q(x)) = \mathbb{E}_{p(x)} \log \frac{p(x)}{q(x)}$$

MLE is Approximating the Real Distribution

$$\arg\max_{\theta} \mathbb{E}_{x \sim p_{data}(x)} \log p_{model}(x;\theta)$$

What is the optimal p_{model} ?

MLE is equivalent to

$$\arg\min_{\theta} D_{\mathrm{KL}}(p_{data}(x) || p_{model}(x;\theta))$$

$D_{KL} \geq 0$ is a distance metric between two distributions, it is 0 when the two distributions are identical

$$D_{\mathrm{KL}}(p(x)||q(x)) = \mathbb{E}_{p(x)} \log \frac{p(x)}{q(x)}$$

When data is all the data from the world, then MLE is learning a distribution for the world

$$KL(P \| Q) = E_P\left(\log \frac{P}{Q}\right) \qquad \log P - \log Q$$

$$\arg\min_\theta \; KL\left[P_{data}(x) \| P_{model}(x;\theta)\right]$$

$$= \arg\min_\theta \left[E_{P_{data}(x)}\left(\log P_{data}(x) - \log P_{model}(x;\theta)\right)\right]$$

$$= \arg\min_\theta \; -E_{P_{data}(x)} \log P_{model}(x;\theta) = \arg\max_\theta E_{P_{data}(x)} \log P_{model}$$

MLE

KL divergence

$$KL(P\|Q) = E_P \log \frac{P}{Q} \geq 0$$

Jensen Inequality

$= 0$ when $P(x) = Q(x)$

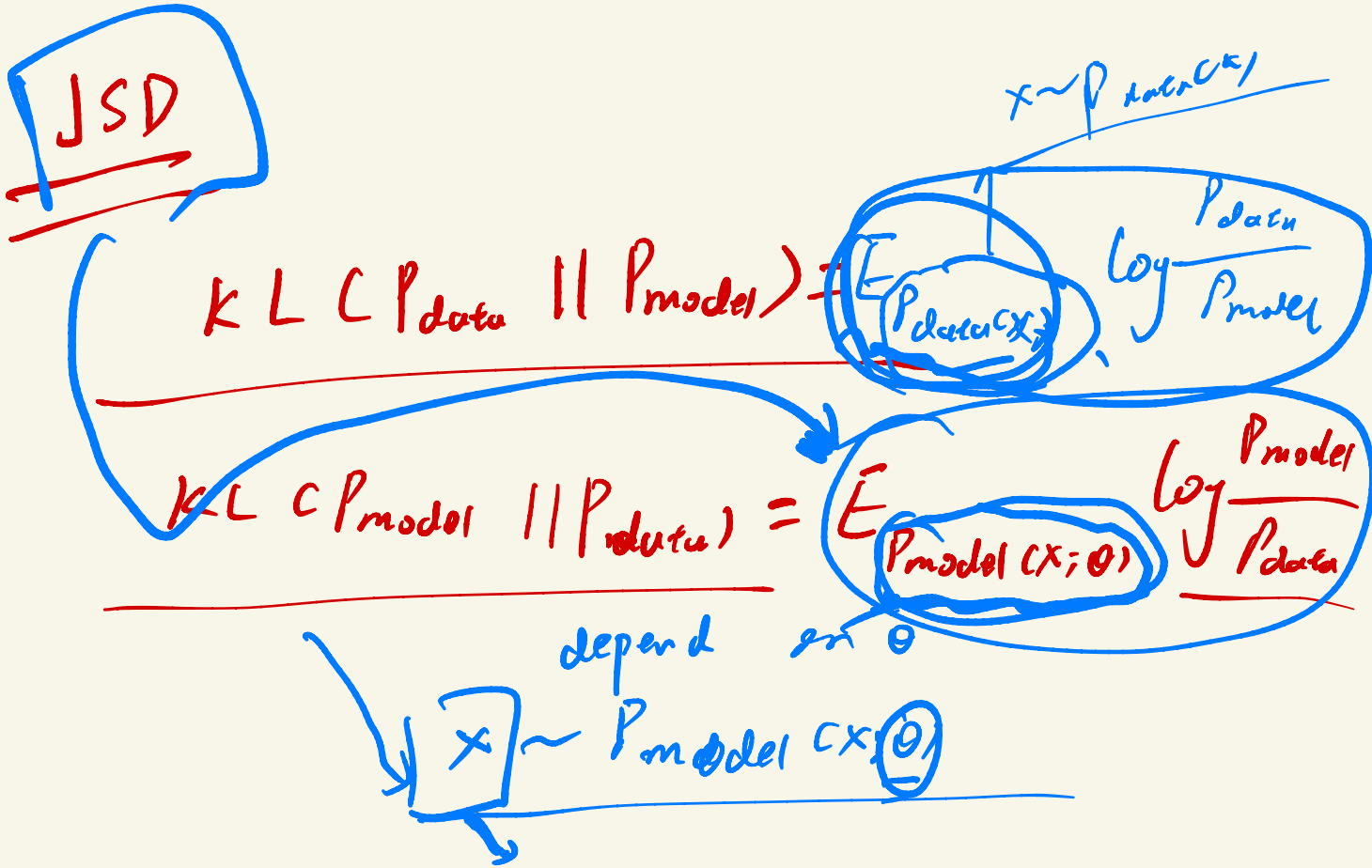
$$-E_P \log \frac{Q}{P} \geq -\log E_P \frac{Q}{P}$$

concave

$$E_P \frac{Q}{P} = \int_x P(x) \cdot \frac{Q(x)}{P(x)}$$

$Q(x) = P(x)$ everywhere $= 1$

$=$


JSD
$x \sim p_{data}(x)$
$KL(p_{data} \| p_{model}) = E_{p_{data}(x)} \log \frac{p_{data}}{p_{model}}$
$KL(p_{model} \| p_{data}) = E_{p_{model}(x;\theta)} \log \frac{p_{model}}{p_{data}}$
depend on θ
$x \sim p_{model}(x;\theta)$

Biased/Unbiased Estimator

Biased/Unbiased Estimator

Suppose we want to estimate a true quantity θ^*, and our estimation is $\hat{\theta}$, then we define the bias of the estimation as:

Biased/Unbiased Estimator

Suppose we want to estimate a true quantity θ^*, and our estimation is $\hat{\theta}$, then we define the bias of the estimation as:

$$bias = \mathbb{E}(\hat{\theta}) - \theta^*$$

Biased/Unbiased Estimator

Suppose we want to estimate a true quantity θ^*, and our estimation is $\hat{\theta}$, then we define the bias of the estimation as:

$$bias = \mathbb{E}(\hat{\theta}) - \theta^*$$

When does the estimation converges to the true value when we have infinite data samples?

Biased/Unbiased Estimator

Suppose we want to estimate a true quantity θ^*, and our estimation is $\hat{\theta}$, then we define the bias of the estimation as:

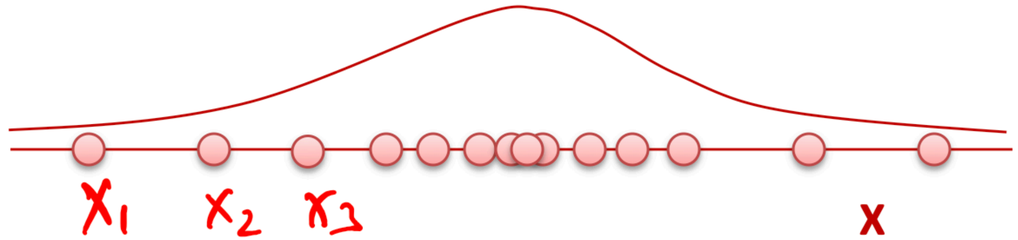
$$bias = \mathbb{E}(\hat{\theta}) - \theta^*$$

When does the estimation converges to the true value when we have infinite data samples?

$bias \to 0$, $Var(\hat{\theta}) \to 0$

Learn Parameters from Data with MLE

Data, D =



Approximate the mean and variance of the data

Data are **i.i.d.**:

- **Independent** events
- **Identically distributed** according to Gaussian distribution

MLE for Gaussian Mean and Variance

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

MLE for Gaussian Mean and Variance

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

Are the estimations biased?

MLE for Gaussian Mean and Variance

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Are the estimations biased?

Unbiased estimator: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$

Max A Posterior (MAP) Estimation

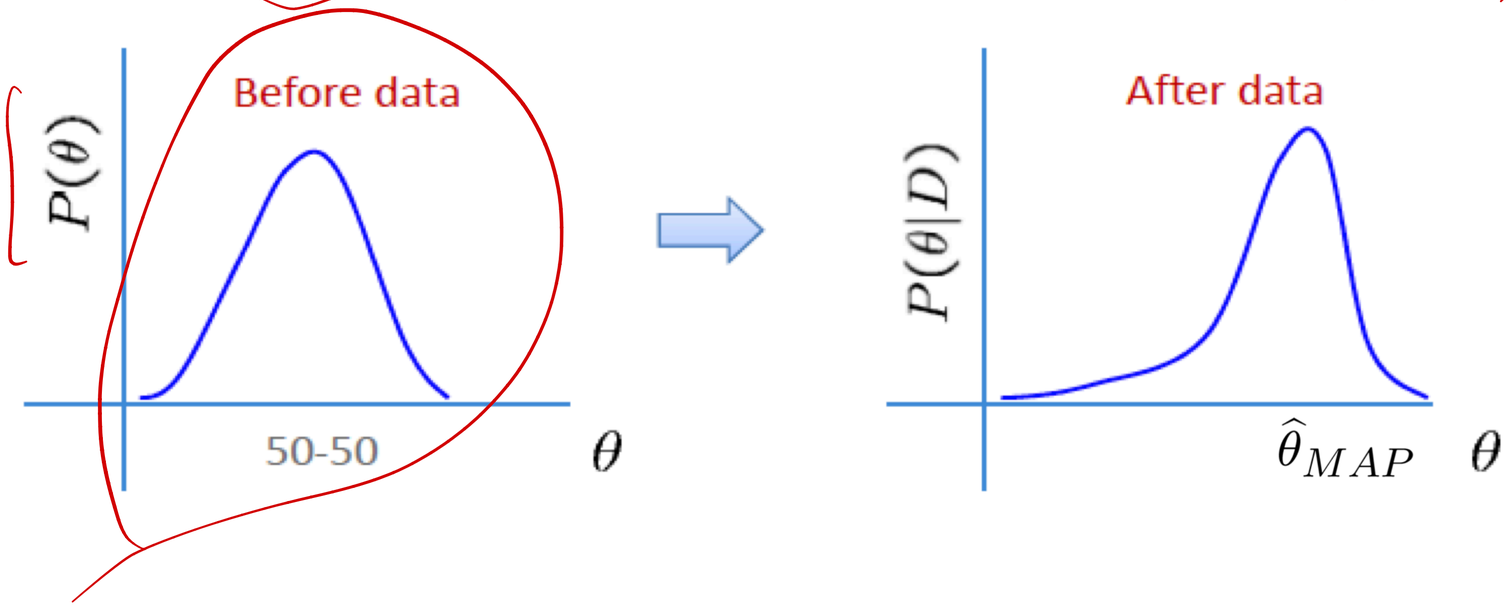
Max A Posterior (MAP) Estimation

Bring prior knowledge to the parameter, define the prior $P(\theta)$. The posterior distribution is $P(\theta \mid D)$. D is the training dataset

$\theta \sim N(0,1)$

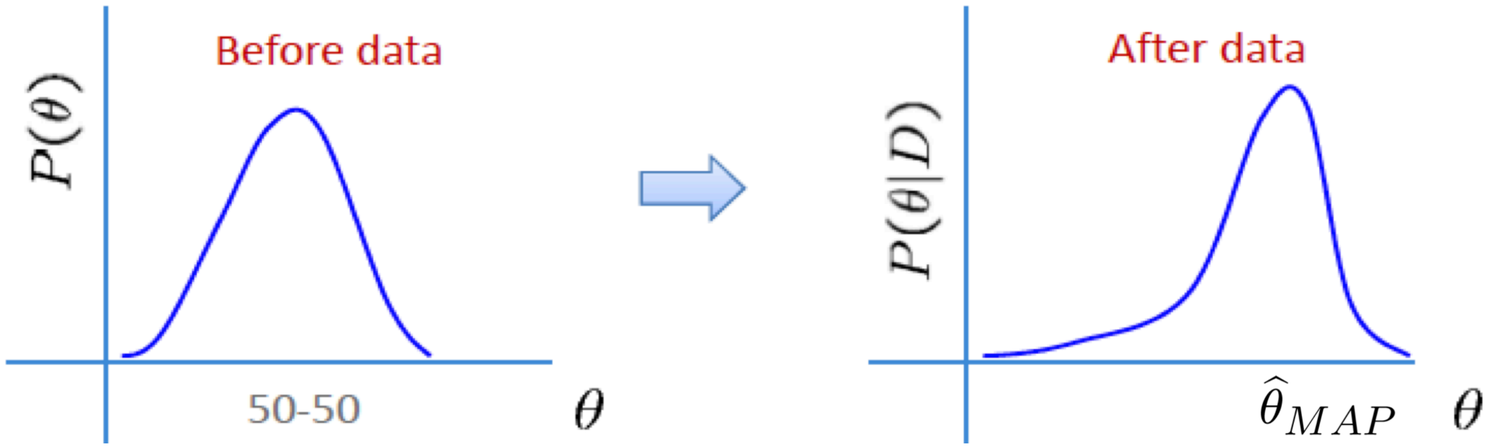
Max A Posterior (MAP) Estimation

Bring prior knowledge to the parameter, define the prior $P(\theta)$. The posterior distribution is $P(\theta \mid D)$. D is the training dataset



Max A Posterior (MAP) Estimation

Bring prior knowledge to the parameter, define the prior $P(\theta)$. The posterior distribution is $P(\theta \mid D)$. D is the training dataset

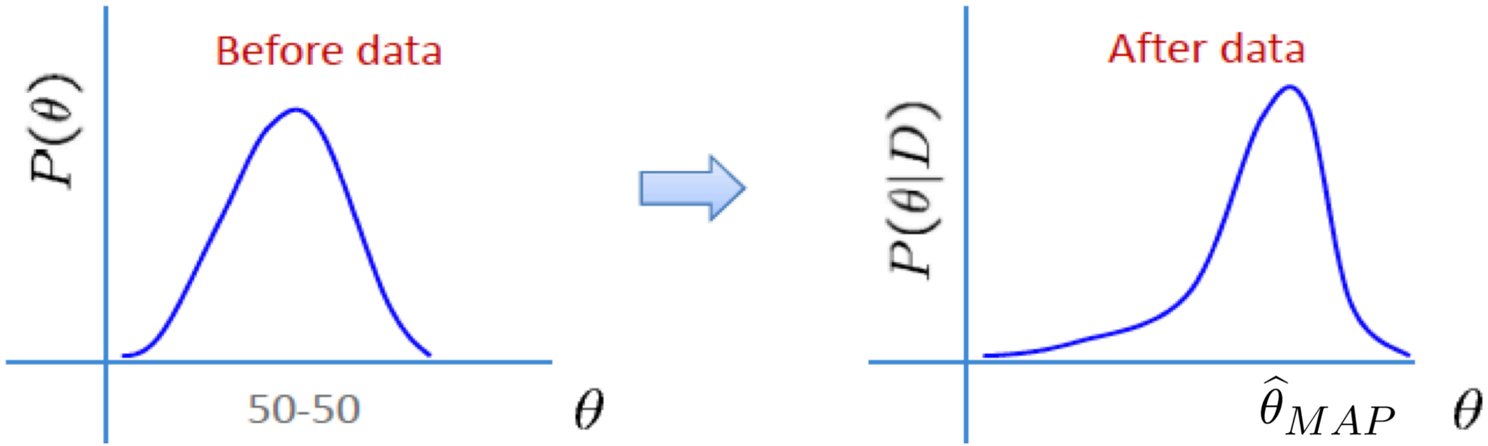


$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$$

$$= \arg\max_{\theta} P(D \mid \theta) P(\theta)$$

Max A Posterior (MAP) Estimation

Bring prior knowledge to the parameter, define the prior $P(\theta)$. The posterior distribution is $P(\theta \mid D)$. D is the training dataset

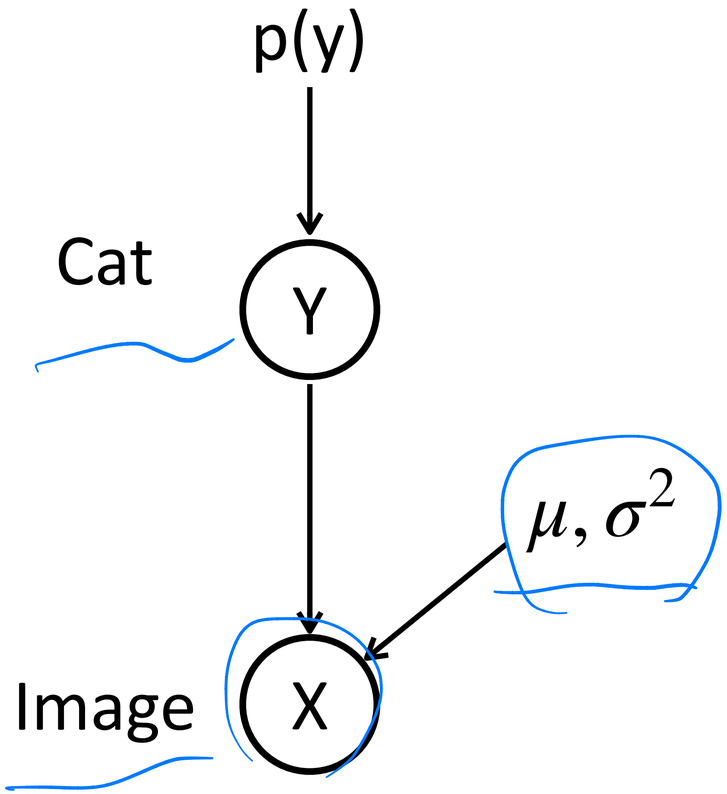


$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$$
$$= \arg\max_{\theta} P(D \mid \theta) P(\theta)$$

Bayesian statistics: there is no “parameters” in the world, all are posterior distributions to estimate

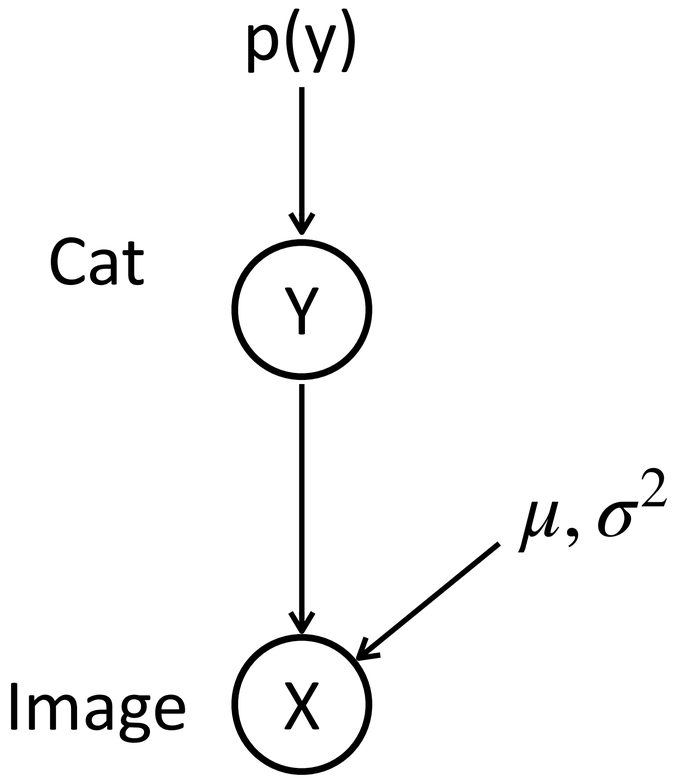
Max A Posterior (MAP) Estimation

Max A Posterior (MAP) Estimation

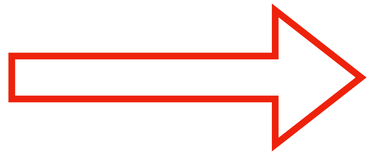


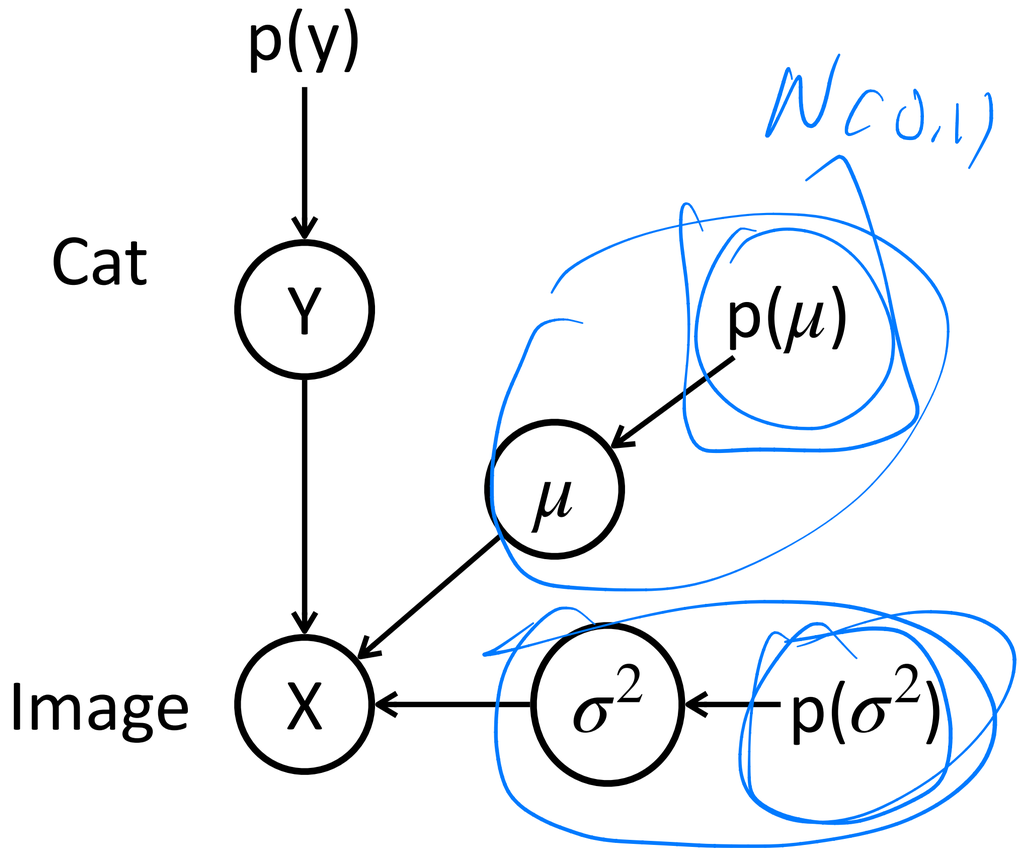
Frequentist

Max A Posterior (MAP) Estimation



Frequentist





Bayesian

How to Choose Prior

- Inject prior human knowledge to regularize the estimate
  - Could learn better if data is limited

How to Choose Prior

- Inject prior human knowledge to regularize the estimate
  - Could learn better if data is limited

- Posterior easy to compute
  - Conjugate prior

exponential family (PDF)

Conjugate Prior

Conjugate Prior

If $P(\theta)$ is conjugate prior for $P(D|\theta)$, then Posterior has same form as prior

Posterior = Likelihood x Prior

$$P(\theta|D) = P(D|\theta) \times P(\theta)$$

Conjugate Prior

If $P(\theta)$ is conjugate prior for $P(D|\theta)$, then Posterior has same form as prior

Posterior = Likelihood x Prior

$P(\theta|D) = P(D|\theta) \times P(\theta)$

P(theta)	P(D|theta)	P(theta|D)
Gaussian	Gaussian	Gaussian
Beta	Bernoulli	Beta
Dirichlet	Multinomial	Dirichlet

MLE vs. MAP

Maximum Likelihood estimation (MLE)

Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

Maximum *a posteriori* (MAP) estimation

Choose value that is most probable given observed data and prior belief

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

When are they the same?

Thank You!
Q & A