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

COMP 5212
Machine Learning
Lecture 9

Naive Bayes, MLE, MAP

Junxian He
Oct 8, 2024

Recap: Generative Models

Recap: Generative Models

X

$p(y \mid x)$

Discriminative

Cat Y

Recap: Generative Models

$\boxed{P(x, y)}$

X

$p(y\,|\,x)$ Discriminative

Cat Y

Cat Y $p(y)$

Generative $p(x\,|\,y)$

X

Recap: Generative Models

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

P(y) P(x|y)

P(y|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: Generative Models Compared to Discriminative Models

Pros:

- Generative models can generate data (generation, data augmentation)
- Inject prior information through the prior distribution $P(y)$
- May be learned in an unsupervised way when y is not available
- Modeling data distribution is a fundamental goal in AI

apple

Cons:

- Often underperforms discriminative models on discriminative tasks because of stronger assumptions on the data

Naive Bayes

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

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

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?

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 $x_j = 0$

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 $x_j = 0$

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

dictionary

count

order

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 $x_j = 0$

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

Dimension is the size of the dictionary

50000

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 $x_j = 0$

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

vocabulary

Dimension is the size of the dictionary

Email Spam Classification

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

$\underbrace{P(y)}_{\text{binary}}$ $\underbrace{P(x|y)}$

Email Spam Classification

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

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

2^{50000} possibilities

$P(x, y)$

$\boxed{P(x \mid y=0)}$ $\forall x$

$P(x \mid y=1)$ $\forall x$

$2^{50000} \times 2$ parameters

Email Spam Classification

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

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

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

Email Spam Classification

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

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

$P(x_i \mid y=1) \neq P(x_i \mid y=1, x_j)$

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

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

$P(x \mid y)$ $\quad$ $P(\quad \mid y=1$

x_i = whether "computer" appears $\quad$ x_j = whether "software" appears

Email Spam Classification

Email Spam Classification

chain rule

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

conditional independence

Email Spam Classification

Autoregressive

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

Email Spam Classification

Autoregressive

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

Parameters

$p(y=0) = 1-$

$p(y=1$

$$\underbrace{\phi_{j|y=1} = p(x_j = 1 \,|\, y = 1)}_{50000}, \quad \underbrace{\phi_{j|y=1} = p(x_j = 1 \,|\, y = 0)}_{50000}, \quad \phi_y = \underbrace{p(y = 1)}_{1}$$

Email Spam Classification

shared NN

$x_1, x_2 \cdots x_k, y$

Autoregressive

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

Parameters

$2 \times 2^{50000} + 1$

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

50000 x 2 + 1 parameters (dict size is 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

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

j

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

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

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

n: # all emails

Maximum Likelihood Estimation

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

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

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

Prediction

Prediction

$P(y)$ $P(x|y)$

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

Prediction

$$
\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x)} \\
&= \underbrace{\frac{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{d} p(x_j|y=1)\right) p(y=1) + \left(\prod_{j=1}^{d} p(x_j|y=0)\right) p(y=0)}}_{\text{no need}}
\end{aligned}
$$

Naive Classifier

$$P(y=1|X) \propto P(X|y=1)P(y=1)$$

$$P(y=0|X) \propto P(X|y=0)P(y=0)$$

same denominator

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?

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

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

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

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

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

Laplace Smoothing

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

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

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

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

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

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

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

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

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

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

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

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 $\{z^{(1)}, \cdots, z^{(n)}\}$

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

j

$= 0$

Laplace Smoothing

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

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

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

Laplace Smoothing

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

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

$$\Downarrow$$

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

Why adding k to the denominator?

$$\boxed{\sum_{j=1}^{k} \phi_j = 1}$$

Laplace Smoothing

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

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

$$\Downarrow$$

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

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)

Maximum Likelihood Estimation (MLE)

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

Maximum Likelihood Estimation (MLE)

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

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

$p_{data}(x)$

intractable

Maximum Likelihood Estimation (MLE)

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

MLE

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

In practice:

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

Maximum Likelihood Estimation (MLE)

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

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

In practice:

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

$x^{(i)}$ 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

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

In practice:

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

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

Monte Carlo Estimation of Expectation

$$\underbrace{E_{x \sim q(x)} f(x)}$$

$q(x)$: easy to sample from

expectation cannot incompeatde

unbiased estimator

$\Downarrow$ approximate

$$\underbrace{\frac{1}{N} \sum_{i=1}^{N} f(x_i)}$$

$x_i \sim q(x)$

$N \longrightarrow$

Maximum Likelihood Estimation (MLE)

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

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

In practice:

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

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

Monte Carlo Estimation of Expectation

Why can we make this approximation?

central limit ~~the~~ ~~theo~~ theorem

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

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

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

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

Monte Carlo Estimation of Expectation

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

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

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

Monte Carlo Estimation of Expectation

p(x)

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

easy?

In practice, n is often small, like 1 sample

n=1

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

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

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

$p(x) = f(x)$ x_0

example:

$p(z) \sim N(0,1)$

$z \to x$

$x = f_{NN}(z)$

GANs

$z \sim N(0,1)$ $x = f(z)$ $[x_0]$

$p(x) = \int_z p(z)\, p(x|z)$

Sampling and Evaluation of Distributions

$P(x)$

- Some distributions are easy to sample from but hard to compute the probability value (hard to evaluate) implicit distribution [GANs
  - 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 $P(x) \propto f(x)$ → x is continuous
  - How to sample from a distribution efficiently is a separate topic

explicit generative models

$P(x)$

MLE is Approximating the Real Distribution

MLE

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

relationship?

$p_{model_\theta}(x;\theta) \rightarrow p_{data}(x)$

fundamental goal

MLE is Approximating the Real Distribution

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

What is the optimal p_{model} ?

MLE is Approximating the Real Distribution

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

What is the optimal p_{model} ?

MLE is equivalent to

KL divergence

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

MLE is Approximating the Real Distribution

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

What is the optimal p_{model} ?

equivalent

MLE is equivalent to

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

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

MLE is Approximating the Real Distribution

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

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

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

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

$$\arg\min_{\theta} D_{KL}(P_{data}(x)\,||\,P_{\theta}(x))$$

constant

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

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

$$= \arg\max_{\theta} E_{P_{data}(x)} \log P_{\theta}(x) \quad \text{MLE}$$

MLE is Approximating the Real Distribution

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

What is the optimal p_{model}? $\rightarrow$ p_{data}

$JSD = KL(p(x) \| q(x)) + KL(q(x) \| p(x))$

MLE is equivalent to

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

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

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

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

$$\underset{\theta}{\arg\min}\ d(P_{data}, P_{model_\theta})$$

$d(\cdot, \cdot)$ distance, KL is only one option

when $d = KL(\ ) \Rightarrow$ MLE

$d = JSD(\ ) \Rightarrow$ GAN

GAN

$$KL(q\|p) = \underbrace{E_{q(x)} \log \frac{q(x)}{p(x)}}$$ distance

not symmetric

$$KL(q\|p) \neq KL(p\|q)$$

$$d = KL(q\|p) + KL(p\|q)$$ symmetric

$$d \geq 0$$

Biased/Unbiased Estimator

Biased/Unbiased Estimator

Suppose we want to estimate a true quantity $\theta*$, 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 $\theta*$, 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 $\theta*$, 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 $\theta*$, 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?

$$\underbrace{bias \rightarrow 0}_{\substack{\nRightarrow \\ \text{unbiased}}}, \quad \underbrace{Var(\hat{\theta}) \rightarrow 0}$$

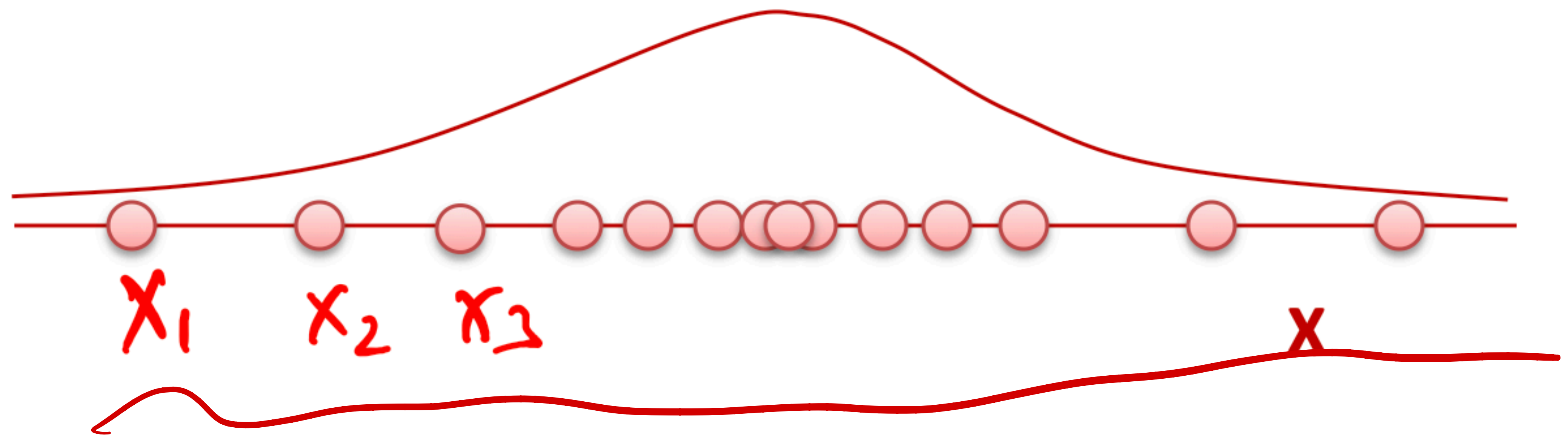
bias = 0 unbiased

bias ≠ 0, bias → 0

bias = $\frac{1}{n}$

Learn Parameters from Data with MLE

Data, D =



Approximate the mean and variance of the data

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

- **Independent** events

μ, σ^2

- **Identically distributed** according to Gaussian distribution

MLE for Gaussian Mean and Variance

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

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

$Var = E((X - E(X))^2)$

MLE for Gaussian Mean and Variance

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

unbiased

$$E[\hat{\mu}_{MLE}] = \frac{\sum_{i=1}^{n} E(x_i)}{n} = \frac{\sum_{i=1}^{n}\mu}{n} = \mu$$

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

$$E[\hat{\sigma}^2_{MLE}] \neq \sigma^2$$

Are the estimations biased?

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?

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

Max A Posterior (MAP) Estimation

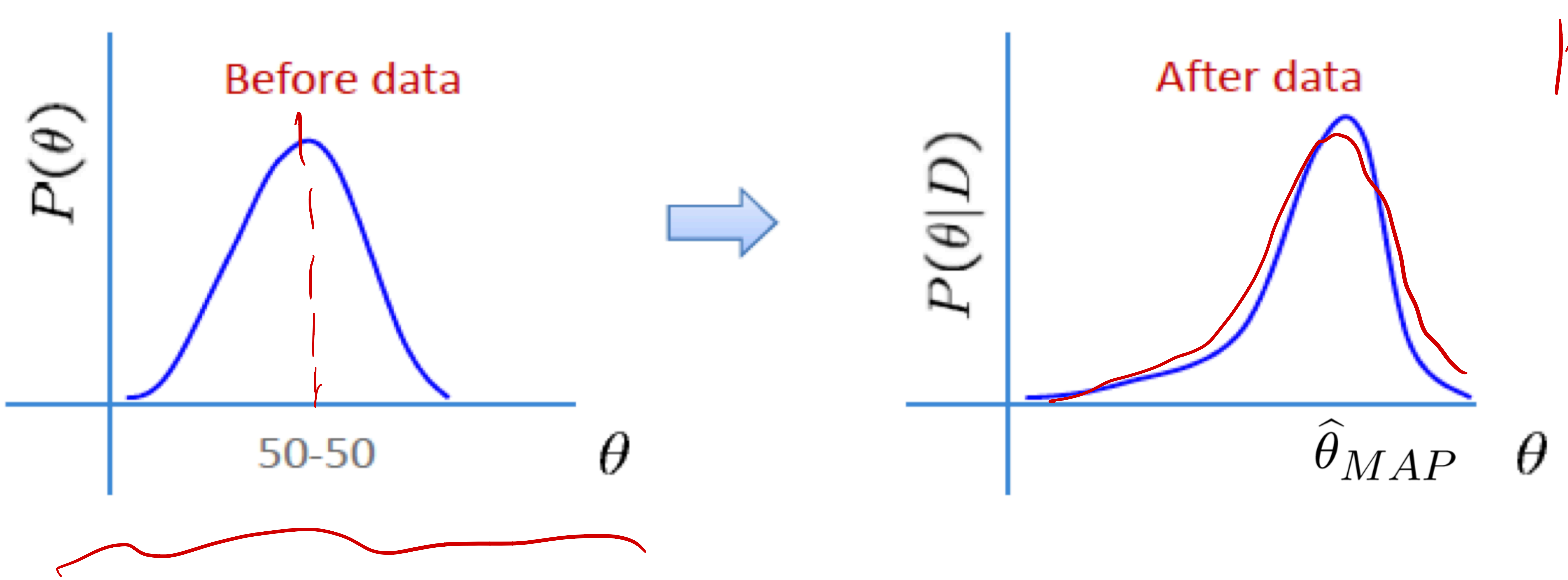
Max A Posterior (MAP) Estimation

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

$P(\theta)$ $P(\theta \mid x)$

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

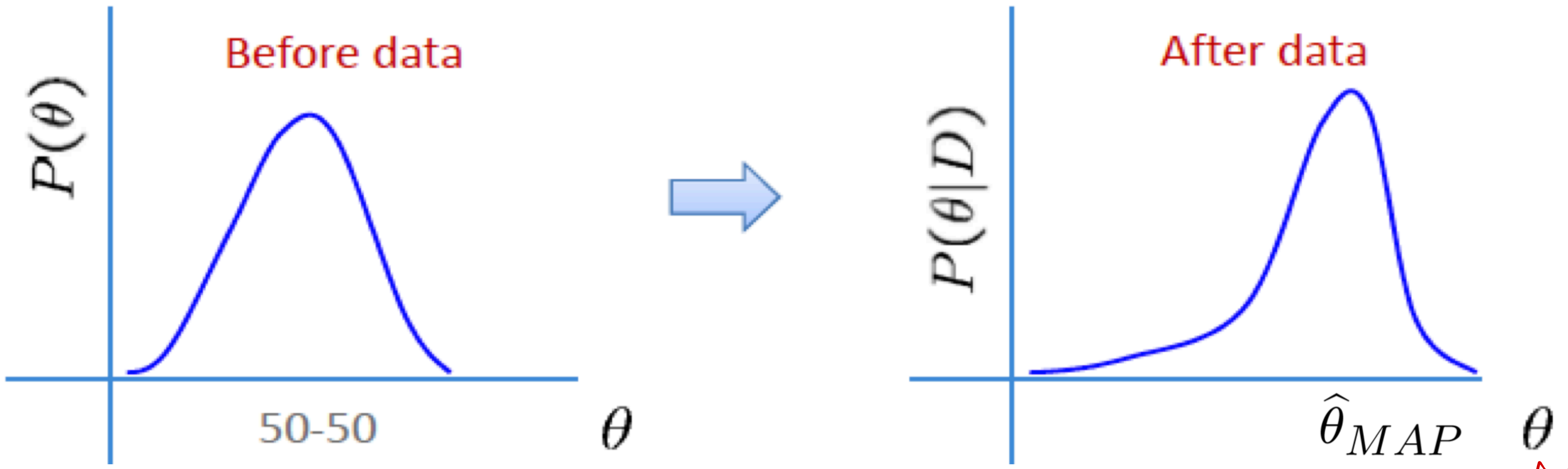


$P(\theta|x)$

θ

Max A Posterior (MAP) Estimation

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



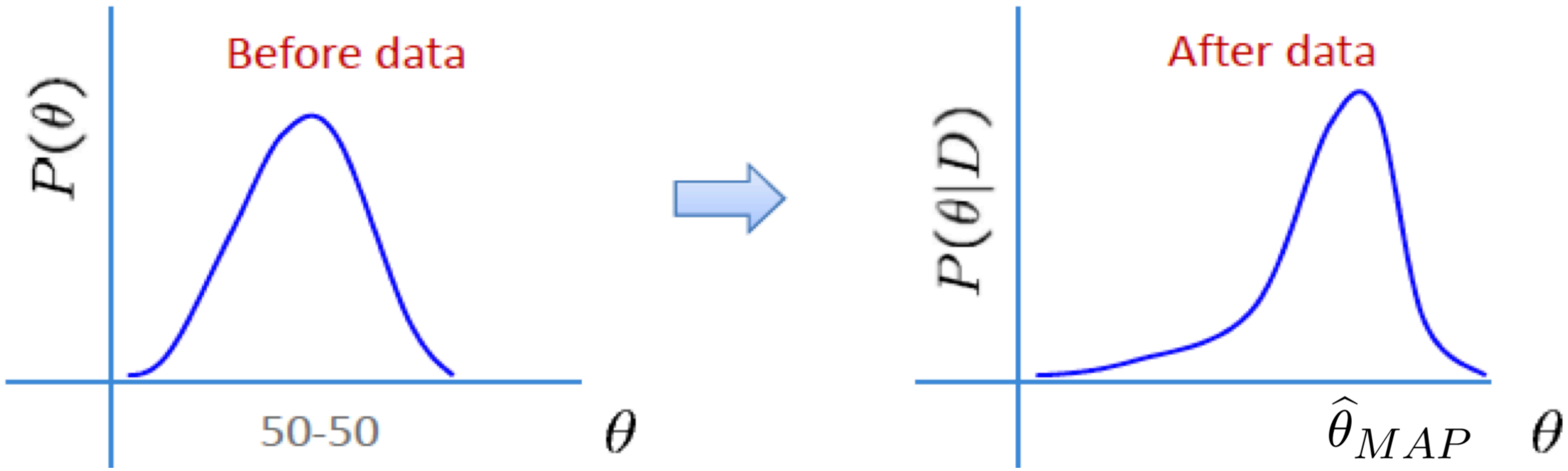
$$
\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\theta} \; P(\theta \mid D) \\
&= \arg\max_{\theta} \; P(D \mid \theta) P(\theta)
\end{aligned}
$$

$\arg\max_{\theta} P(D|\theta)$

MLE

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

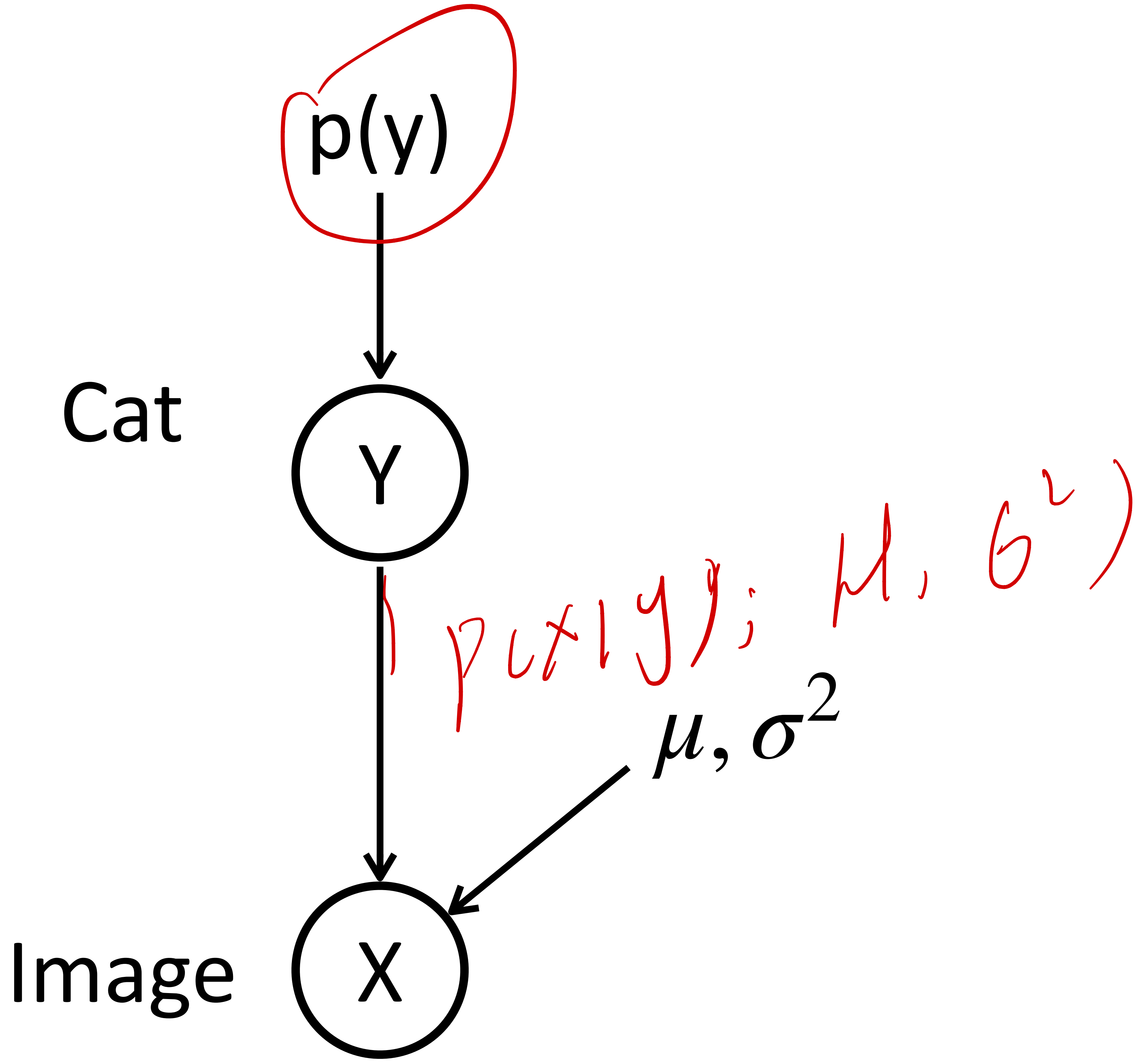


$$
\begin{aligned}
\widehat{\theta}_{MAP} &= \arg\max_{\theta} \; P(\theta \mid D) \\
&= \arg\max_{\theta} \; P(D \mid \theta) P(\theta)
\end{aligned}
$$

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

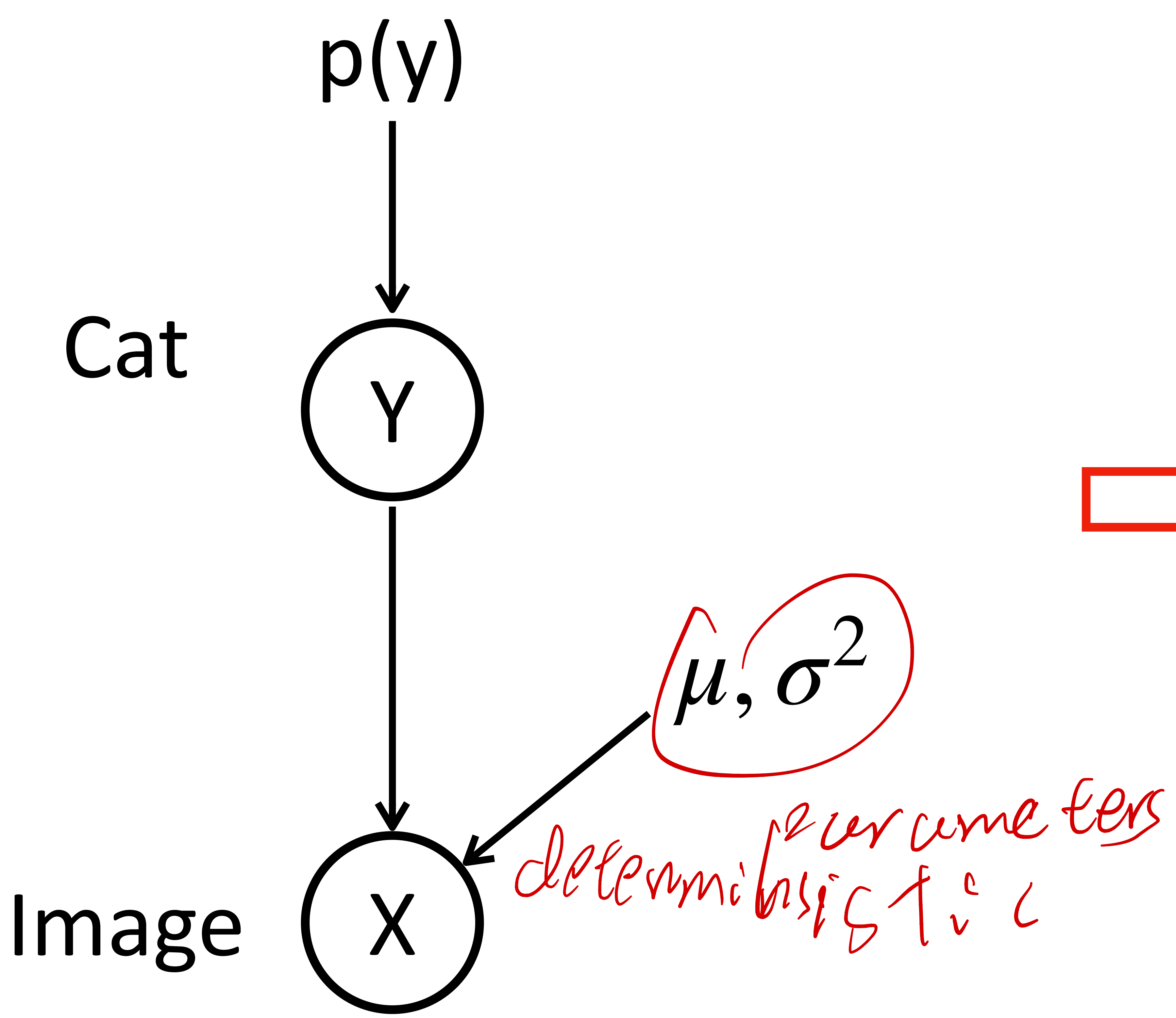
Max A Posterior (MAP) Estimation

Max A Posterior (MAP) Estimation



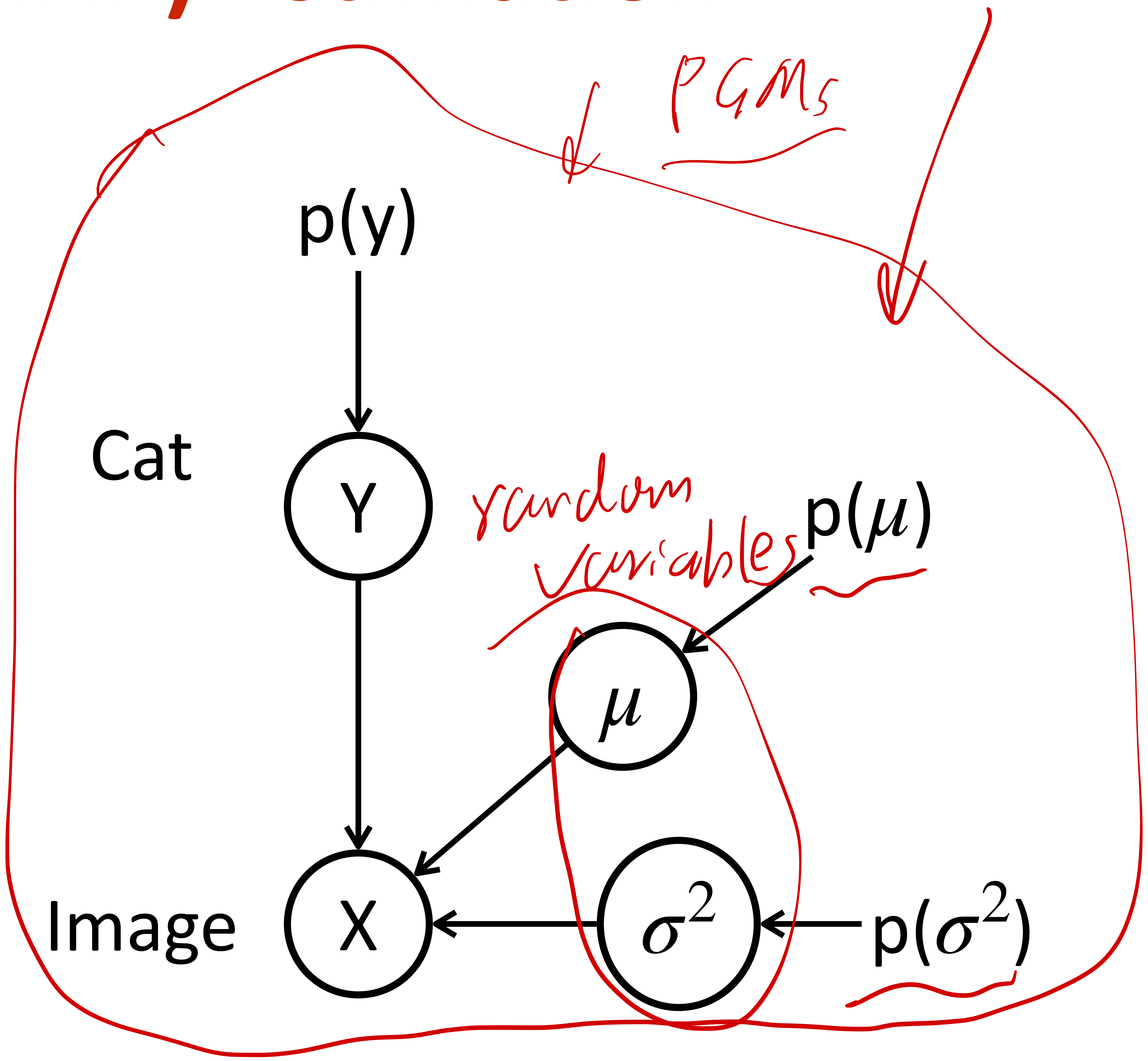
Frequentist

Max A Posterior (MAP) Estimation



Frequentist

Bayesian NN



Bayesian

$P(\mu \mid X)$ $P(\sigma^2 \mid X)$

How to Choose Prior

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

$P(y)$ z continuous

$P(z) \sim \mathcal{N}(0,1)$

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

$P(z) \to P(x|z)$

$P(z|x)$ same family as $P(z)$

$P(z)$

z

$P(x|z)$

x

$\arg\max_\theta \log P_\theta(x)$

Expectation Maximization

$E_{P(z|x)} f(x,z)$

$z \sim P(z|x)$

$P(z|x) \propto P(z) P(x|z)$

Conjugate Prior

Conjugate Prior

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

$$\text{Posterior} = \text{Likelihood} \times \text{Prior}$$

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

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

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

Maximum *a posteriori* (MAP) estimation

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

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

When are they the same?

Thank You!
Q & A