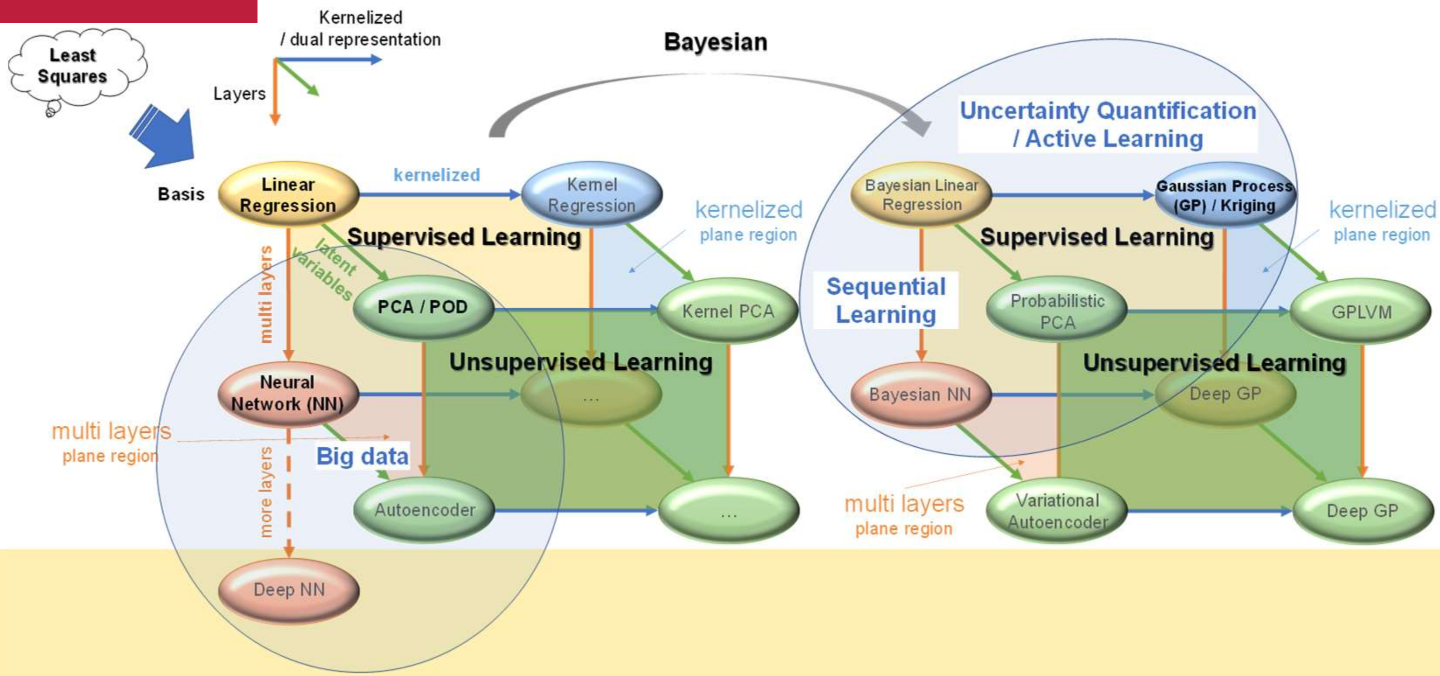


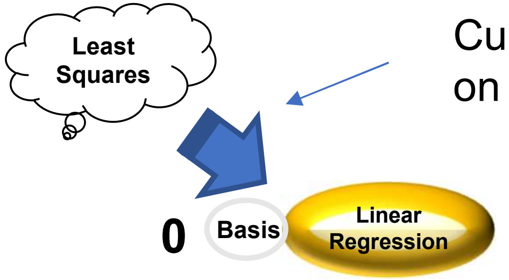
Scientific Machine Learning

Lecture 3: Curve-Fitting Revisiting / Probability Distribution

Dr. Daigo Maruyama

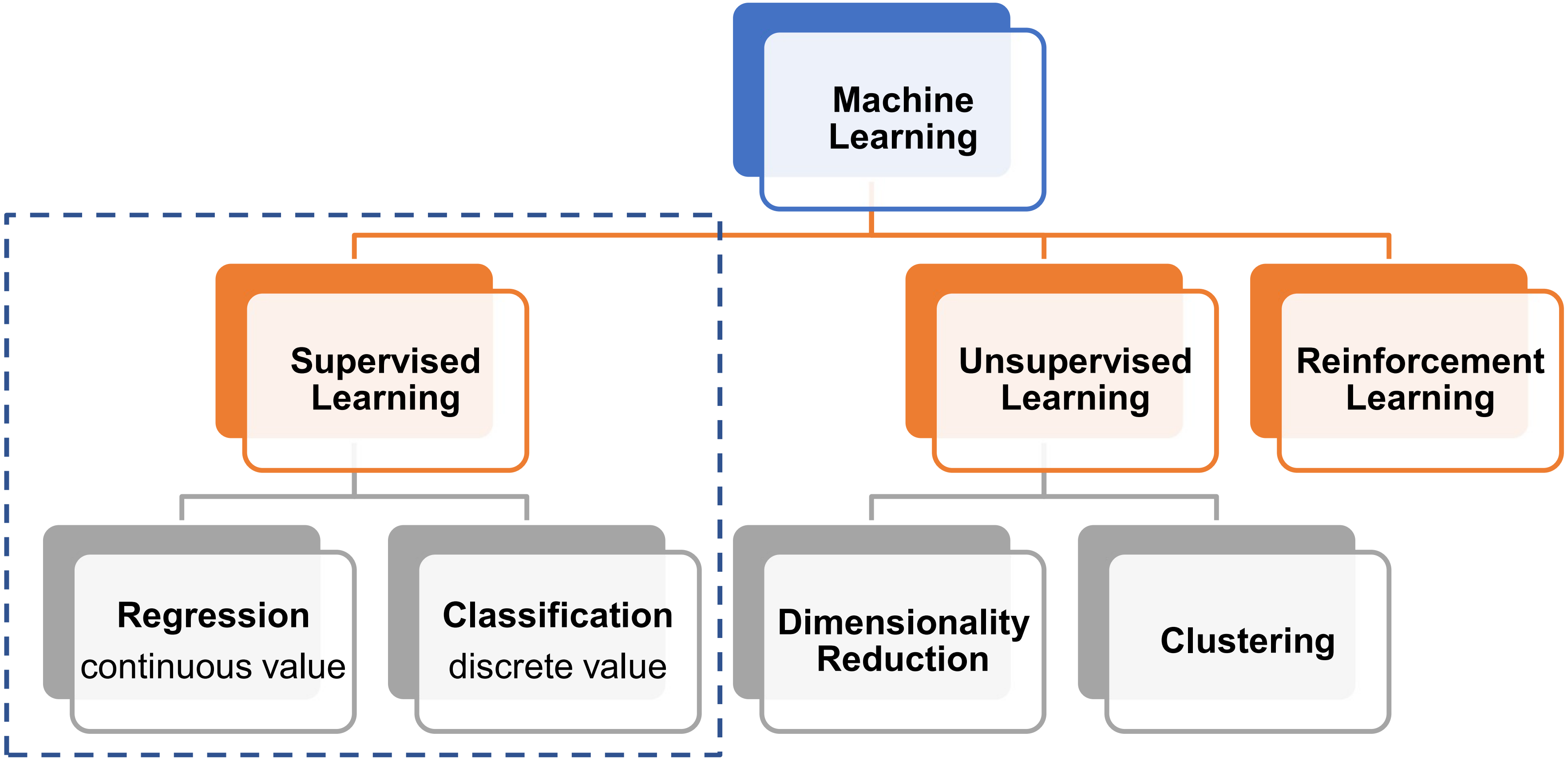
Prof. Dr. Ali Elham

Key Components (Current Position)



Curve-Fitting Revisiting:
on the way to Linear Regression

Machine Learning Classification by Use/Application - Revisit



In this course, machine learning classification is done by **methods and their concepts**.

➡ Then the use/application is naturally derived/understood.

Lecture content

- Maximum Likelihood Estimation (continued from Lecture 2)
- Curve Fitting Revisiting
- Probability Distributions

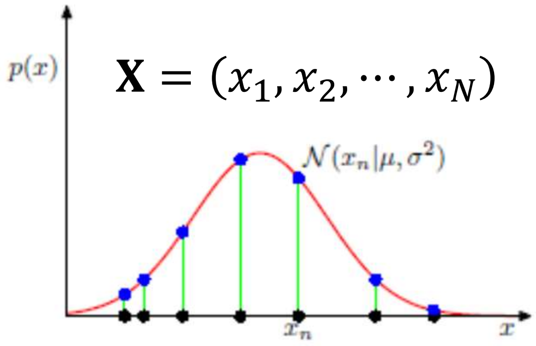
The lecture of this time basically follows the 1st and 2nd chapters of the book:
Christopher M. Bishop "Pattern Recognition And Machine Learning" Springer-Verlag (2006)
The name of this book is shown as “PRML” when it is referred in the slides.

Lecture content

0. Maximum Likelihood Estimation
(continued from Lecture 2)

Likelihood Function - Review

Likelihood function: a probability of data



Data points are assumed to be generated from <u>a distribution</u> (pdf) $p(x)(= p(x|\mu, \sigma))$.

1. **<u>Independent</u> and identically distributed (i.i.d.)**

$$p(x_1, x_2) = p(x_1)p(x_2) = \prod_{i=1}^{2} p(x_i)$$

2. $p(x_i|\mu, \sigma)$:
the probability when the data point x_i is generated from the distribution $p(x|\mu, \sigma)$.

➡ We can define the probability when all the data points are generated from the distribution $p(x|\mu, \sigma)$, which is $p(\mathbf{X}|\mu, \sigma)$.

a probability of the data $\mathbf{X}$

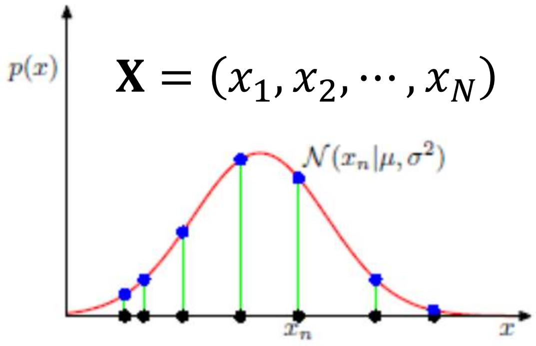
$$\underline{p(\mathbf{X}|\mu, \sigma)} = \prod_{i=1}^{N} p(x_i|\mu, \sigma)$$

When this probability is regarded as a function of the parameters μ and σ, $p(\mathbf{X}|\mu, \sigma)$ is not a probability anymore.

But useful for estimation of the parameters μ, σ!

Maximum Likelihood Estimation (MLE)

Likelihood function: a probability of data



$$\mathbf{X} = (x_1, x_2, \cdots, x_N)$$

a probability of the data X

$$p(\mathbf{X}|\mu,\sigma) = \prod_{i=1}^{N} p(x_n|\mu,\sigma)$$

$$L(\mu,\sigma) \equiv -\ln p(\mathbf{X}|\mu,\sigma) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 + \frac{N}{2}\ln\sigma^2 + \frac{N}{2}\ln 2\pi$$

Take negative log

Maximum Likelihood Estimation (MLE)

$$\hat{\boldsymbol{w}}, \hat{\sigma} = \underset{\boldsymbol{w},\sigma}{\operatorname{argmax}}\, p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\sigma)$$

Maximizing the likelihood function with respect to the parameters μ and σ

Optimization problem

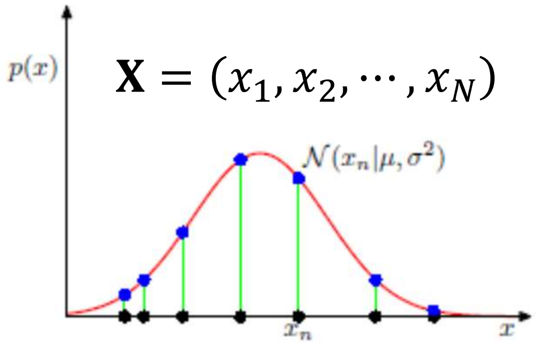
$$\frac{\partial L(\mu,\sigma)}{\partial\mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)$$

when $\frac{\partial L(\mu,\sigma)}{\partial\mu} = 0$,

$$\begin{cases} \hat{\mu} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} x_n \\ \hat{\sigma}^2 = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(x_n-\hat{\mu})^2 \end{cases}$$

Maximum Likelihood Estimation (MLE)

Likelihood function: a probability of data



a probability of the data $\mathbf{X}$

$$p(\mathbf{X}|\mu, \sigma)$$

Maximum Likelihood Estimation (MLE)

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}}\, p(\mathbf{X}|\mu, \sigma)$$

Maximize:
- the probability of the data given the parameters: $p(\mathbf{X}|\mu, \sigma)$ Likelihood
- the probability of the parameters given the data: $p(\mu, \sigma|\mathbf{X})$ Posterior

Which is correct?

Bayes' theorem

under certain conditions:

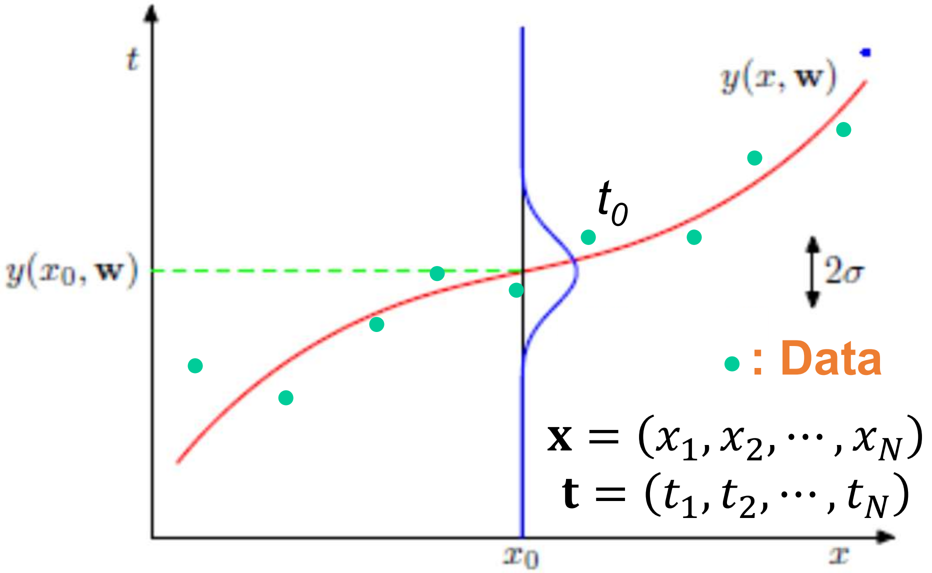
$$p(\mathbf{X}|\mu, \sigma) \propto p(\mu, \sigma|\mathbf{X})$$

Lecture content

1. Curve Fitting Revisiting

Curve Fitting Revisiting

The least square method and the regularization method are summarized in perspectives based in the **probability theory**.



based on PRML, p. 29

x: deterministic variable
t: random variable

Consider a pdf of the output t_0 at a given input x_0

$$p(t_0|x_0)$$

We **assume** that this pdf is a Gaussian distribution parametrized by μ and σ.

Probabilistic model

$$p(t_0|x_0, \mu, \sigma) = \mathcal{N}(t_0|\mu, \sigma^2)$$

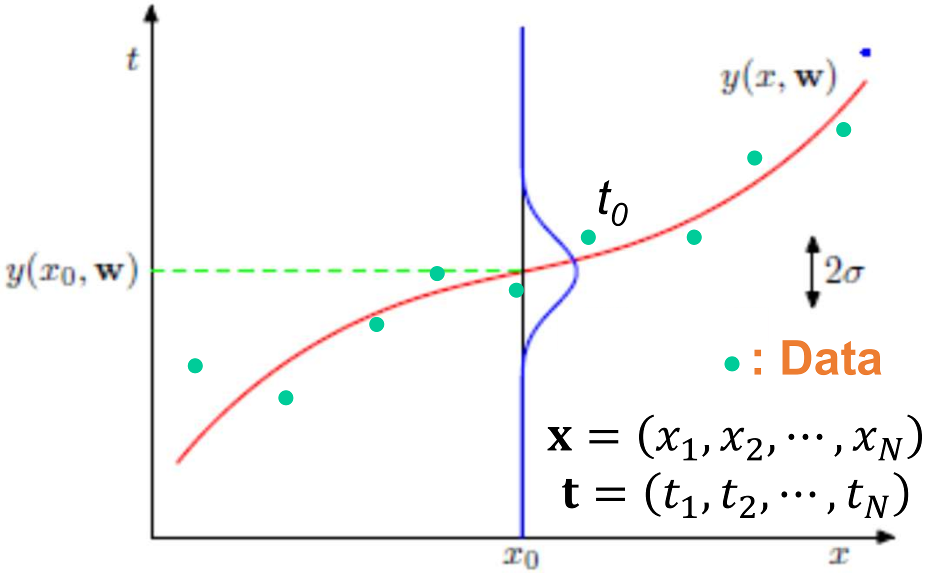
We further assume that the mean μ is a function the input x.

$$\mu = y(x, \boldsymbol{w})$$

e.g. $y(x, \boldsymbol{w})$ is a polynomial function.

Curve Fitting Revisiting

The least square method and the regularization method are summarized in perspectives based in the **probability theory**.



based on PRML, p. 29

The probabilistic model is now (for arbitrary input x):

$$p(t|x, \boldsymbol{w}, \sigma) = \mathcal{N}(t|y(x, \boldsymbol{w}), \sigma^2)$$

Consider parameters μ, σ when they make the probability of data $\mathbf{X}$ (**Likelihood function**) maximum.

Consider the likelihood function.

$$p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \sigma) = \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, \boldsymbol{w}), \sigma^2)$$

x: deterministic variable
t: random variable

Maximum Likelihood Estimation (MLE)

$$\widehat{\boldsymbol{w}}, \hat{\sigma} = \underset{\boldsymbol{w}, \sigma}{\operatorname{argmax}}\, p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \sigma)$$

Please confirm that $\widehat{\boldsymbol{w}}$ by MLE is identical to that by the least square method.

Curve Fitting Revisiting

Likelihood function

$$p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \sigma) = \prod_{i=1}^{N} \mathcal{N}(t_i | y(x_i, \boldsymbol{w}), \sigma^2)$$

$$\hat{\boldsymbol{w}}, \hat{\sigma} = \underset{\boldsymbol{w},\sigma}{argmax}\, p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \sigma)$$

Take negative log

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{\{t_i - y(x_i, \boldsymbol{w})\}^2}{2\sigma^2}\right]$$

$$L(\boldsymbol{w}, \sigma) \equiv -\ln p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \sigma) = \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{N} \{t_i - y(x_i, \boldsymbol{w})\}^2}_{\text{the least square term}} + \frac{N}{2} \ln(2\pi\sigma^2)$$

Minimizing $L(\boldsymbol{w}, \sigma)$ w.r.t. $\boldsymbol{w}$ leads to the least square method.

- Numerical errors in computing the likelihood function can be eased by taking the log.
- The negative log of likelihood function is normally called Error Function.

Curve Fitting Revisiting

Bayes' theorem

Likelihood **Prior**

$$\underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\textbf{Posterior}} = \frac{\overbrace{p(\mathcal{D}|\boldsymbol{w})}^{\textbf{Likelihood}}\overbrace{p(\boldsymbol{w})}^{\textbf{Prior}}}{p(\mathcal{D})}$$

Posterior

$\mathcal{D}$: data
$\boldsymbol{w}$: parameters

When we focus on the parameters $\boldsymbol{w}$:

Objective: obtain $\hat{\boldsymbol{w}}$

$$\boxed{p(\boldsymbol{w}|\mathcal{D})} \propto \boxed{p(\mathcal{D}|\boldsymbol{w})}\boxed{p(\boldsymbol{w})}$$

obtained

set as prior of $\boldsymbol{w}$

similar concept to the regularization (constraint on $\boldsymbol{w}$)

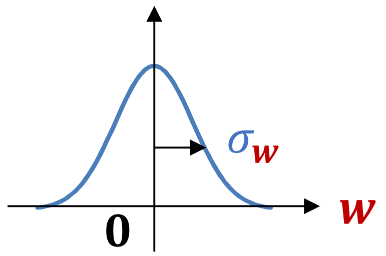
the new function to be optimized

Curve Fitting Revisiting

e.g. $p(\boldsymbol{w})$: a Gaussian distribution around $\mathbf{0}$

$$p(\boldsymbol{w}|\sigma_{\boldsymbol{w}}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, {\sigma_{\boldsymbol{w}}}^2\mathbf{I})$$

$$p(\boldsymbol{w}|\sigma_{\boldsymbol{w}}) = \frac{1}{\left(\sqrt{2\pi{\sigma_{\boldsymbol{w}}}^2}\right)^{M+1}} exp\left[-\frac{\|\boldsymbol{w}\|^2}{2{\sigma_{\boldsymbol{w}}}^2}\right]$$



Prior distribution of $\boldsymbol{w}$

Bayes' theorem $p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})$

$$-\ln p(\boldsymbol{w}|\mathcal{D}) \propto -\ln p(\mathcal{D}|\boldsymbol{w}) - \ln p(\boldsymbol{w})$$

$$= \frac{1}{2\sigma^2}\sum_{i=1}^{N}\{t_i - y(x_i, \boldsymbol{w})\}^2 + \frac{N}{2}\ln(2\pi\sigma^2) + \frac{1}{2{\sigma_{\boldsymbol{w}}}^2}\|\boldsymbol{w}\|^2$$

$$-\ln p(\boldsymbol{w}|\mathcal{D}) \propto \sum_{i=1}^{N}\{t_i - y(x_i, \boldsymbol{w})\}^2 + \left(\frac{\sigma}{\sigma_{\boldsymbol{w}}}\right)^2\|\boldsymbol{w}\|^2 = \underbrace{E(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2}$$

where,

$$\lambda \equiv \left(\frac{\sigma}{\sigma_{\boldsymbol{w}}}\right)^2$$

The same function as that for the regularization

Lecture content

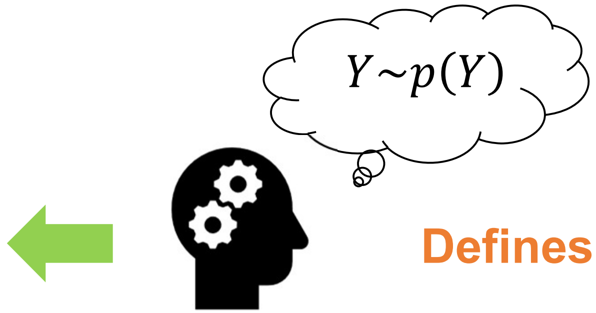
2. Probability Distributions

Machine Learning Modeling (Revisit)

Probabilistic model is hypothesis.

Observations

Y

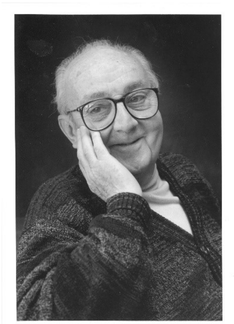


Defines $p(Y)$

We should know some useful models.

"**All models are wrong, but some are useful**."
The aphorism from George Box*

*George E. P. Box "Science and Statistics", *Journal of the American Statistical Association*, 71(791799), 1976.



https://en.wikipedia.org/wiki/George_E._P._Box

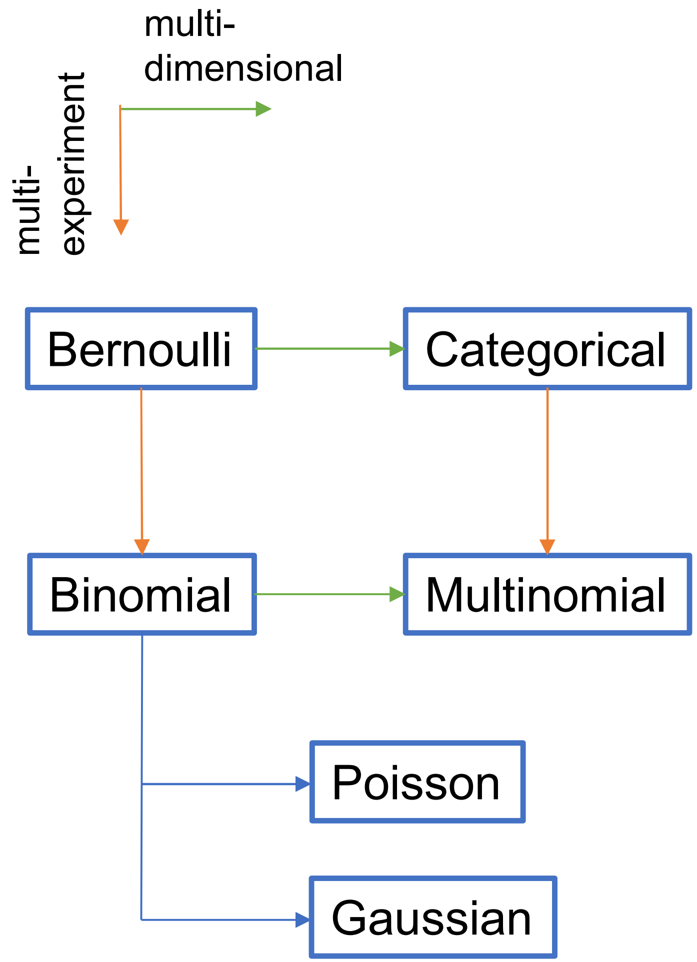
We cannot inquire which is correct, input error or output error.

Probability Distributions

- **Parametric distributions** $p(x|\boldsymbol{\theta})$
  - Discrete probability distributions
    - Bernoulli distribution
    - Binomial distribution
    - Categorical distribution
    - Multinomial distribution
    - Poisson distribution
  - Continuous probability distributions
    - Beta distribution
    - Dirichlet distribution
    - Gaussian distribution
    - Laplace distri

$p(x|\boldsymbol{\theta}) = \mathcal{N}(x|\mu, \sigma^2)$
where, $\boldsymbol{\theta} = (\mu, \sigma)$

- **Non-parametric distributions**



Probability Distributions

- **Parametric distributions** $p(x|\boldsymbol{\theta})$
  - Discrete probability distributions **for Classification / Discrete output**
    - Bernoulli distribution ⬅ 2 classes
    - Binomial distribution
    - Categorical distribution ⬅ multiple classes
    - Multinomial distribution
    - Poisson distribution ⬅ discrete output
  - Continuous probability distributions **for Regression**
    - Beta distribution
    - Dirichlet distribution
    - Gaussian distribution ⬅ almost all cases
    - Laplace distribution
- **Non-parametric distributions**

Probability Distributions

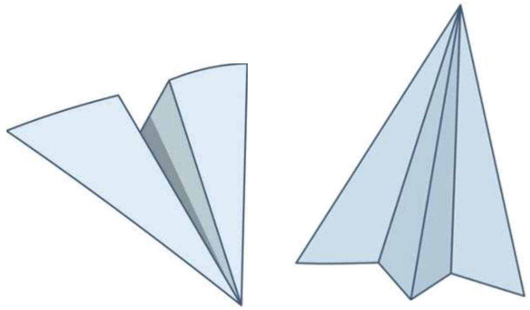
Bernoulli distribution $p(x)$

$$\boxed{\mu^x(1-\mu)^{1-x}}$$

Therefore,

$$\boxed{p(x|\mu) = \mathrm{Bern}(x|\mu)}$$

$x = 1$ $\quad$ $x = 0$ $\quad$ $x \in \{0,1\}$



probability μ $\quad$ probability $1-\mu$ $\quad$ $0 \le \mu \le 1$

Used in Classification (2-classes)

Parameter:
μ: probability of $x = 1$

dataset: $\mathcal{D} = \{x_1, \cdots, x_N\}$

The likelihood

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

$$L(\mu) \equiv -\ln p(\mathcal{D}|\mu) = -\sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

$$\hat{\mu} = \underset{\mu}{\mathrm{argmin}}\, L(\mu)$$

Probability Distributions

Binomial distribution $p(m)$

multiple experiments of Bernoulli

$$\boxed{\binom{N}{m}\mu^m(1-\mu)^{N-m}}$$

Parameters:
μ: probability of $x = 1$ (the same as Bernoulli)
N: number of experiments (can be observations)

m: number of observations of $x = 1$

Therefore,

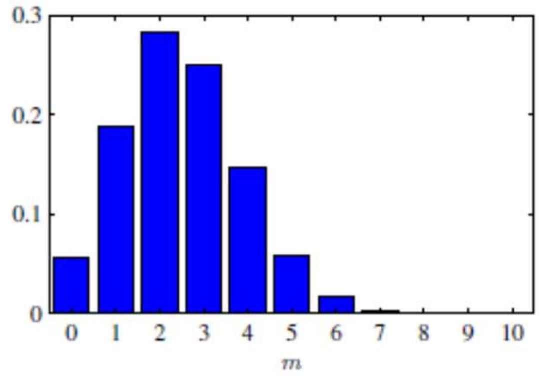
$$\boxed{p(m|N,\mu) = \mathrm{Bin}(m|N,\mu)}$$

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!\,m!} = nCm$$

Basis of:

→ Poisson distribution

→ Gaussian distribution



PRML, p. 70

Please consider $\hat{\mu}$ by MLE when $N = 3$ and $m = 3$.

Probability Distributions

Categorical distribution $p(\boldsymbol{x})$ where, $\boldsymbol{x} = \{x_1, \cdots, x_K\}$, $\sum_{k=1}^{K} x_k = 1$

multi dimensional of Bernoulli

$$\boxed{\prod_{k=1}^{K} {\mu_k}^{x_k}}$$

parameters

$\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)^{\mathrm{T}}$

μ_k: probability of $x_k = 1$

Observations x is now represented as

$K = 6 \qquad \boldsymbol{x} = (0{,}0\underbrace{,1,}_{}0{,}0{,}0)^{\mathrm{T}}$

when 3 is observed

Used in Classification (multiple-classes)

probability theory, wikipedia

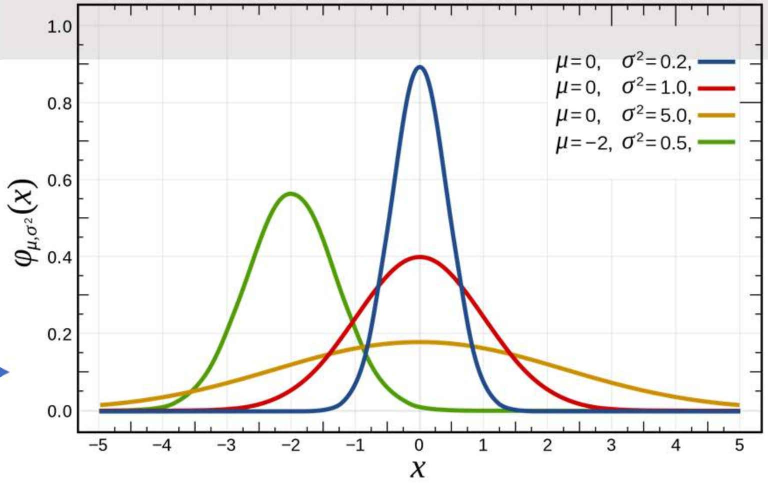
$$\boxed{p(\boldsymbol{x}|\boldsymbol{\mu}) = \mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\mu})}$$

Multinomial distribution

$$\boxed{p(\boldsymbol{m}|\boldsymbol{\mu}, N) = \mathrm{Mult}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{N}) = N! \prod_{k=1}^{K} \frac{{\mu_k}^{m_k}}{m_k!}}$$

Probability Distributions

wikipedia: Gaussian distribution



Binomial distribution

$$p(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

$N \to \infty$
μ: const.

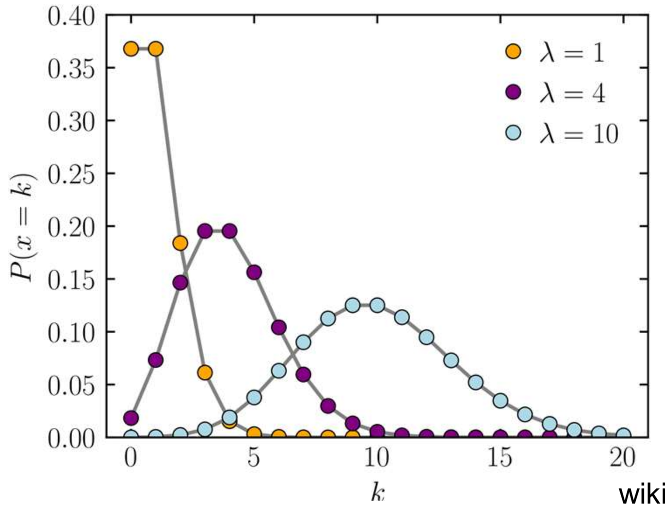
Gaussian distribution

$N \to \infty$
$N\mu$: const.

Poisson distribution

$$\boxed{p(k|\lambda) = \Pr(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}}$$

$\lambda = N\mu$



wikipedia: Poisson distribution

Probability Distributions

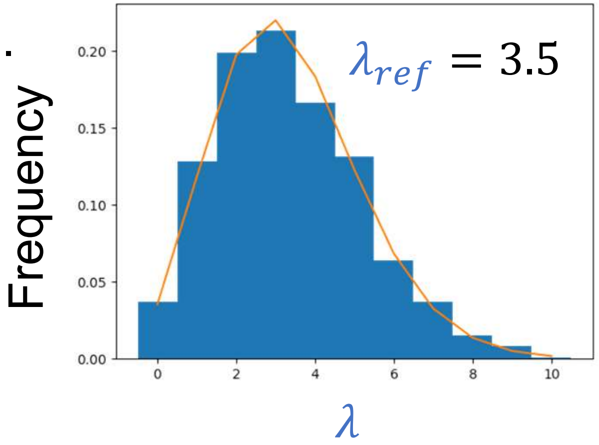
Poisson distribution

$$p(k|\lambda) = \Pr(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

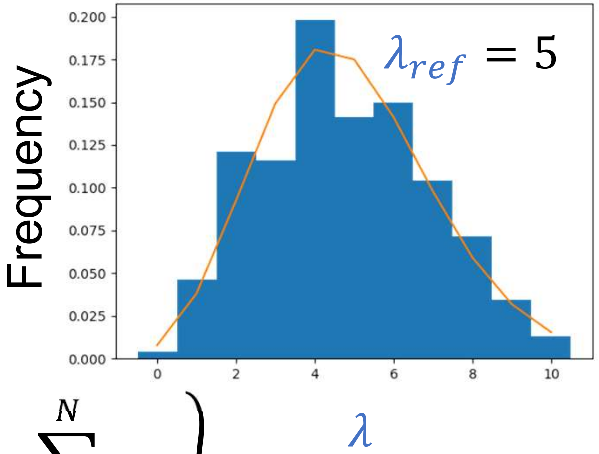
Dataset $\mathbf{k} = \{k_1, \cdots, k_N\}$

$$p(\mathbf{k}|\lambda) = \prod_{n=1}^{N} \Pr(k_n|\lambda)$$

e.g.



by MLE
$\hat{\lambda} = 3.297$



by MLE
$\hat{\lambda} = 4.878$

$$L(\mu) \equiv -\ln p(\mathbf{k}|\lambda) = -\sum_{n=1}^{N} \left\{ k_n \ln\lambda - \lambda - \sum_{n=1}^{N} \ln\lambda \right\}$$

Gaussian Distribution (Normal Distribution)

Carl Friedrich Gauss
(1777-1855)

Born in Braunschweig

Collegium Carolinum at TUBS

Some important topics
related to Gaussian distributions

- Least square method
- Central limit theorem
- Gaussian Process