A Probabilistic Perspective on the Regression and Classification

Qinliang Su （苏勤亮）

Sun Yat-sen University

suqliang@mail.sysu.edu.cn

Outline

- Introduction
- Probabilistic Perspective on Regression
- Probabilistic Perspective on Classification

Perspective from Conditional Probability

- The goal of regression and classification is to predict the *possible output* y given the input data $\boldsymbol{x}$

$$\boldsymbol{x} \xrightarrow{\text{predict}} y$$

- In the regression and classification, the prediction is made by a deterministic function

$$\text{Regression:} \quad f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{w}$$

$$\text{Classification:} \quad f(\boldsymbol{x}) = \sigma(\boldsymbol{x}\boldsymbol{w})$$

From the perspective of probability, to predict the output y given $\boldsymbol{x}$, we just need to model *the conditional probability*

$$p(y|\boldsymbol{x})$$

- With the conditional probability $p(y|\boldsymbol{x})$, the output can be predicted as

$$\text{Mean:} \quad \hat{y} = \int y p(y|\boldsymbol{x}) dy$$

or

$$\text{MAP:} \quad \hat{y} = \arg\max_{y} p(y|\boldsymbol{x})$$

Outline

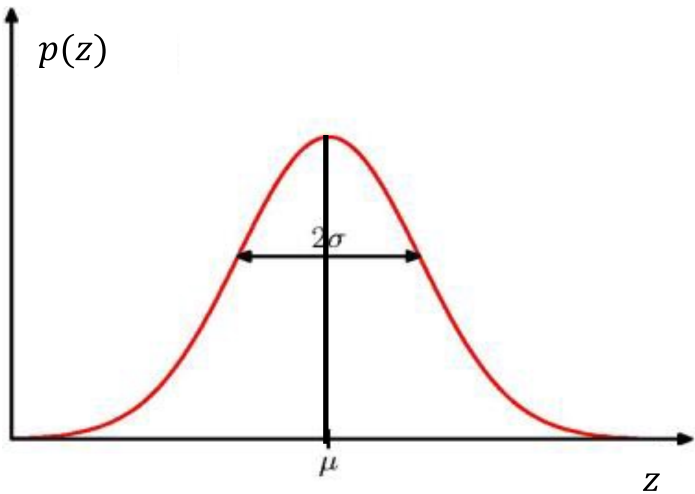
- Introduction
- Probabilistic Perspective on Regression
- Probabilistic Perspective on Classification

Gaussian Distribution

- Univariate Gaussian distribution

$$p(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(z-\mu)^2}{\sigma^2}\right] \triangleq \mathcal{N}(z; \mu, \sigma^2)$$

  - μ is the mean
  - $\sigma^2 = E[(z-\mu)^2]$ is the variance
  - σ is the standard deviation



Bell shape

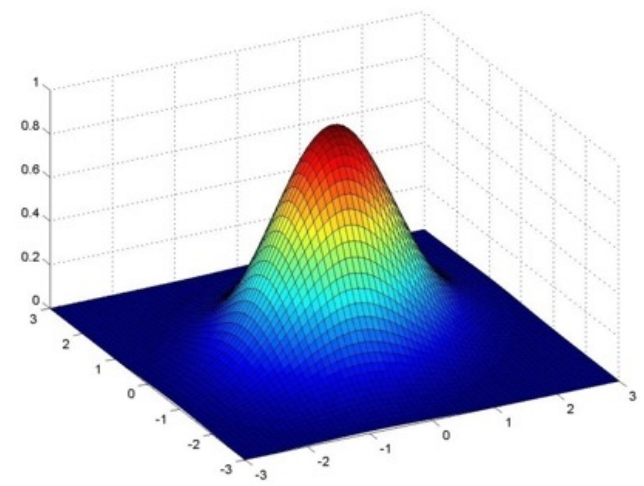
- μ is the *peak* and *center* of the distribution
- σ determines the spread of the distribution

- Multivariate Gaussian distribution

$$p(\boldsymbol{z}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right\} \triangleq \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu},\boldsymbol{\Sigma})$$

  - D is the dimension
  - $\boldsymbol{\mu} \in \mathbb{R}^D$ is the mean vector
  - $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ is the covariance matrix, and $|\boldsymbol{\Sigma}|$ is its determinant

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^T]$$

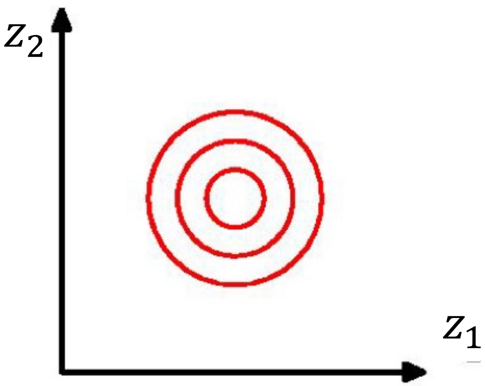


- $\boldsymbol{\mu}$ controls the peak or the central point
- $\boldsymbol{\Sigma}$ controls the shapes of the distribution

- Shapes of the distributions under different kinds of $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

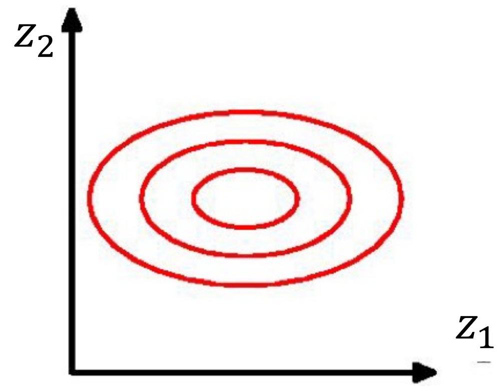
$$\sigma_1^2 = \sigma_2^2$$



(a)

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

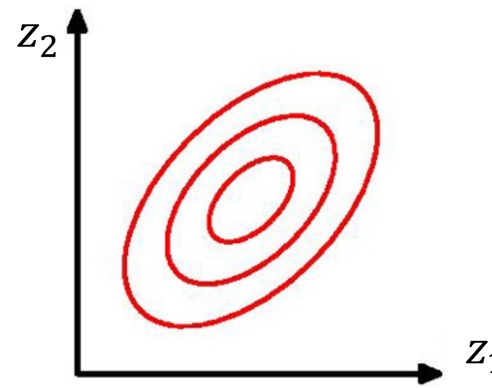
$$\sigma_1^2 > \sigma_2^2$$



(b)

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix}$$

$$\rho \neq 0$$



(c)

- No matter how $\boldsymbol{\Sigma}$ varies, the peak is always located at $\boldsymbol{\mu}$ (unimodal)

- For every covariance matrix $\mathbf{\Sigma}$, it can be decomposed as

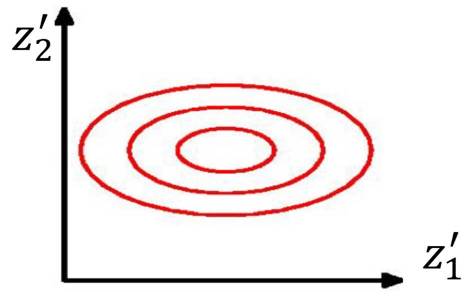
$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$$

  - $\mathbf{U}$ is an orthogonal matrix, with $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$
  - $\mathbf{\Lambda}$ is a *diagonal matrix*

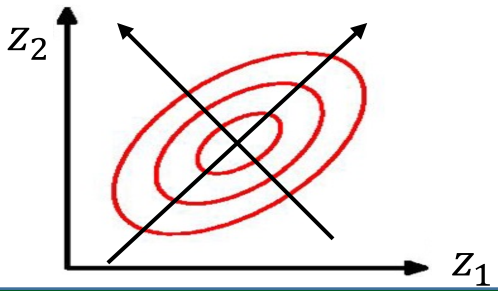
- Letting $\boldsymbol{z}' = \boldsymbol{U}^T\boldsymbol{z}$ and $\boldsymbol{\mu}' = \boldsymbol{U}^T\boldsymbol{\mu}$, the distribution can be expressed as

$$p(\mathbf{z}') = \frac{1}{(2\pi)^{D/2}|\mathbf{\Lambda}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}' - \boldsymbol{\mu}')^T\mathbf{\Lambda}^{-1}(\mathbf{z}' - \boldsymbol{\mu}')\right\}$$

- Thus, the shape of $p(\mathbf{z}')$ looks like



- But the shape of $p(\mathbf{z})$ is rotated by $\boldsymbol{U}$

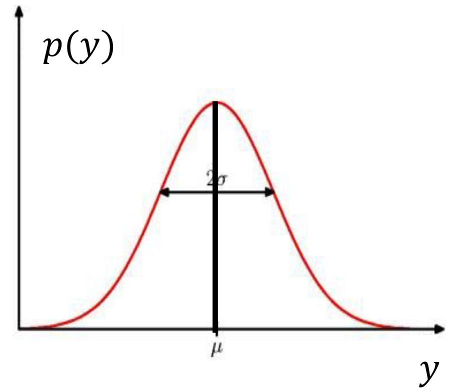


Linear Regression

- From the probabilistic perspective , to make a prediction, we only need to specify the conditional probability distribution $p(y|\boldsymbol{x})$. For regression, we assume the distribution is a normal distribution

$$p(y|\boldsymbol{x};\boldsymbol{w}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{(y-\boldsymbol{x}\boldsymbol{w})^2}{\sigma^2}\right]$$

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



- We make prediction by using the mean of the distribution, *i.e.*,

$$\hat{y} = \boldsymbol{x}\boldsymbol{w}$$

Is $\boldsymbol{w}$ obtained here the same as that in traditional regression?

- The goal of model training is to find the parameter $\boldsymbol{w}$ that maximizes the log-probability, that is,

$$\max_{\boldsymbol{w}} \log p(y|\boldsymbol{x};\boldsymbol{w})$$

Log-likelihood function

- From the expression of $p(y|\boldsymbol{x};\boldsymbol{w})$, we obtain

$$\log p(y|\boldsymbol{x};\boldsymbol{w}) = -\frac{1}{2}\frac{(y-\boldsymbol{x}\boldsymbol{w})^2}{\sigma^2} + constant$$

Thus, maximizing the log-likelihood $\log p(y|\boldsymbol{x};\boldsymbol{w})$ is equivalent to

$$\min_{\boldsymbol{w}} \ (y-\boldsymbol{x}\boldsymbol{w})^2,$$

which is the same as the loss used in the regression

- For N training samples $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$, by assuming they are *i.i.d.*, their joint conditional pdf can be obtained as

$$p\left(y^{(1)}, \cdots, y^{(N)} \middle| \boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{\left(y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{w}\right)^2}{\sigma^2}\right]$$

- The log-likelihood function is

$$\log p\left(y^{(1)}, \cdots, y^{(N)} \middle| \boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\right) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{w}\right)^2 + constant$$

- Maximizing the log-likelihood $\log p\left(y^{(1)}, \cdots, y^{(N)} \middle| \boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\right)$ is equivalent to minimize

$$L(\boldsymbol{w}) = \sum_{i=1}^{N} \left(y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{w}\right)^2,$$

Can you figure out why the summation Σ operator arises here?

which is the same as the loss used in the regression

- From the perspective of probabilistic modelling, linear regression is actually equivalent to

  - Modeling: assuming *conditional distribution to be diagonal Gaussian*

  - Training: training the model by *maximizing the log-likelihood*

Outline

- Introduction
- Probabilistic Perspective on Regression
- Probabilistic Perspective on Classification

Bernoulli Distribution

- The Bernoulli distribution

$$p(z) = \begin{cases} \pi, & \text{if } z = 1 \\ 1 - \pi, & \text{if } z = 0 \end{cases}$$

where $\pi \in [0, 1]$ is the probability of z equal to 1

- The $p(z)$ can be concisely expressed as

$$p(z) = \pi^z \cdot (1 - \pi)^{1-z}$$

where $z = 0$ or 1

Binary Classification

- To achieve binary classification, the conditional probability is assumed to be a Bernoulli distribution

$$p(y|\boldsymbol{x}) = \big(\sigma(\boldsymbol{xw})\big)^{y} \cdot \big(1 - \sigma(\boldsymbol{xw})\big)^{1-y}$$

where $\pi = \sigma(\boldsymbol{xw})$; and $y = 0$ or 1

- The training objective is to maximize the log-likelihood function

$$\log p(y|\boldsymbol{x}) = y \log \sigma(\boldsymbol{xw}) + (1 - y) \log\big(1 - \sigma(\boldsymbol{xw})\big)$$

Recall that the logistic regression minimizes

$$\mathit{cross\ entropy} \triangleq -y \log \sigma(\boldsymbol{xw}) - (1 - y) \log\big(1 - \sigma(\boldsymbol{xw})\big)$$

Maximizing $\log p(y|\boldsymbol{x})$ is equivalent to minimize the cross entropy

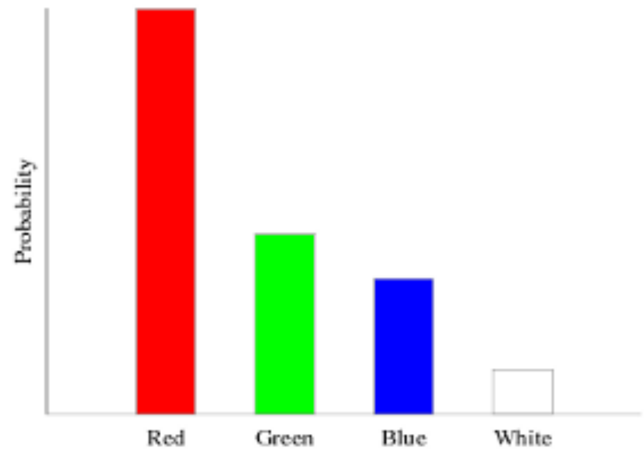
- The logistic regression is equivalent to
    - Modeling: assuming *Bernoulli conditional distribution* for the output
    - Training: training the model by *maximizing the log-likelihood*

Categorical Distribution

- The categorical distribution

$$p(\boldsymbol{z} = onehot_k) = \pi_k$$

  - where $onehot_i = [0, \cdots, 0, 1, 0, \cdots, 0]$ is a vector with the i-th element being the only nonzero element 1
  - $\sum_{k=1}^{K} \pi_k = 1$



- The distribution can be equivalently written as

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

where $\boldsymbol{z}$ is a one-hot vector

Multiclass Classification

- Modeling: By setting the probability π_k as

$$\pi_k = softmax_k(\boldsymbol{x}\boldsymbol{W}),$$

the conditional probability distribution is assumed to obey the categorical distribution

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{k=1}^{K} [softmax_k(\boldsymbol{x}\boldsymbol{W})]^{y_k}$$

- Training: Given a training sample $(\boldsymbol{x}, \boldsymbol{y})$, the model is trained by maximizing the log-likelihood function

$$\log p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{K} y_k \cdot \log\big(softmax_k(\boldsymbol{x}\boldsymbol{W})\big)$$

$$= -\,cross\ entropy$$

Summary

- The regression, logistic and multi-class regressions can be placed under one common framework

1) Modeling: assume different conditional pdfs for the outputs y

    - Regression: *Gaussian distribution*
    - Logistic regression: *Bernoulli distribution*
    - Multiclass logistic regression: *Categorical distribution*

2) Training: maximize the log-likelihood functions