



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 x

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

- In the previous regression and classification, the prediction is given by some deterministic functions

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

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

From the perspective of probability, to predict the output y given x , we just need to model *the conditional probability*

$$p(y|x)$$

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

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

or

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

Outline

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

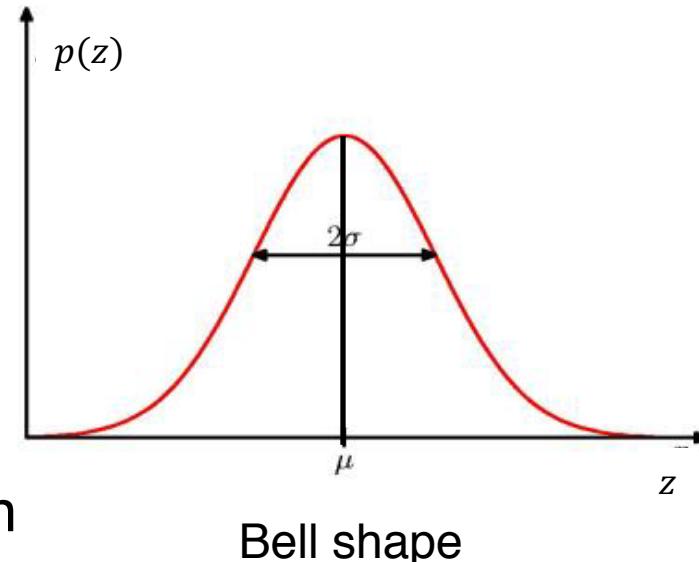
Gaussian Distribution

- 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

- μ is the *peak* and *central* of the distribution
- σ determine the spread of the distribution

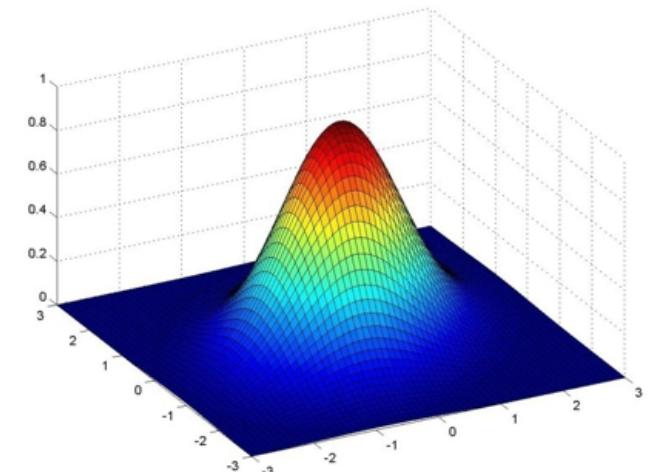


- Multivariate Gaussian distribution

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

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

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

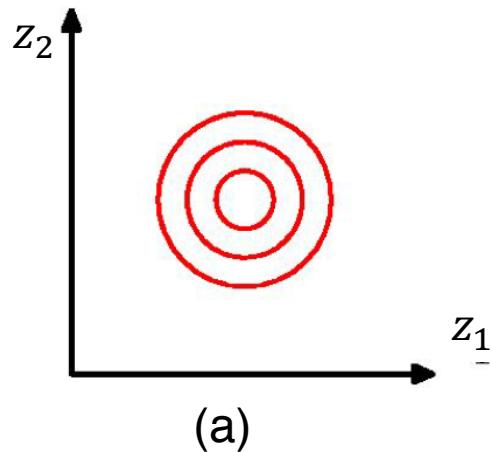


- $\boldsymbol{\mu}$ controls the peak or the central point
- Σ controls the shapes of the distribution

- Shapes under different kinds of Σ

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

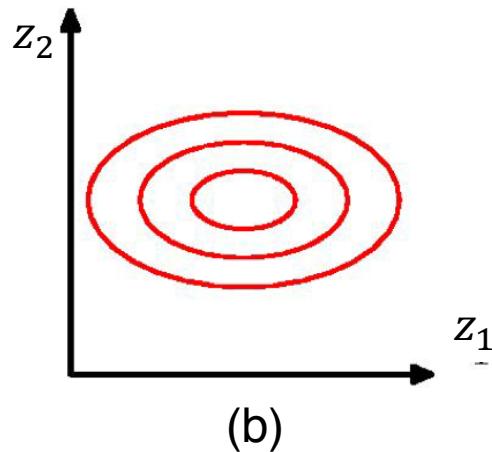
$$\sigma_1^2 = \sigma_2^2$$



(a)

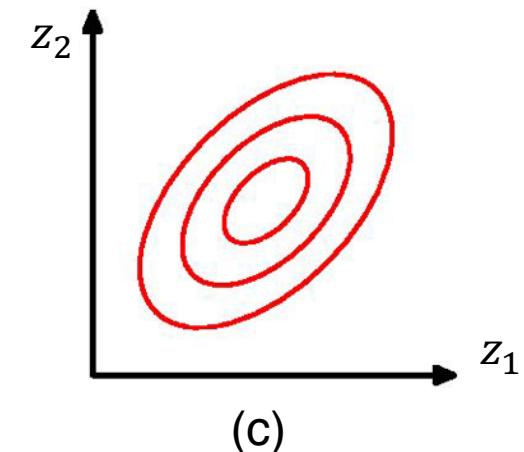
$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

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



(b)

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



(c)

- No matter how Σ varies, the peak is always located at μ (unimodal)

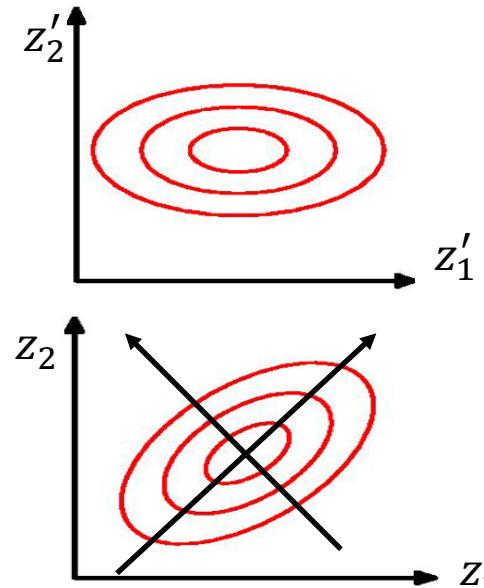
- For every covariance matrix Σ , it can be decomposed as

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

- \mathbf{U} is an orthogonal matrix, with $\mathbf{U}\mathbf{U}^T = \mathbf{I}$
- Λ is a *diagonal matrix*
- By letting $\mathbf{z}' = \mathbf{U}\mathbf{z}$ and $\boldsymbol{\mu}' = \mathbf{U}\boldsymbol{\mu}$, the distribution can be expressed as

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

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

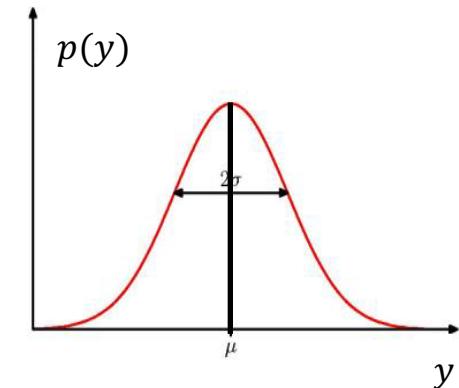


- But the shape of $p(\mathbf{z})$ is rotated \mathbf{U}

Linear Regression

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

$$\begin{aligned} p(y|x; \mathbf{w}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y - \mathbf{x}\mathbf{w})^2}{\sigma^2}\right] \\ &= \mathcal{N}(y; \mathbf{x}\mathbf{w}, \sigma^2) \end{aligned}$$



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

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

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

- Training the model aims to find the parameter \mathbf{w} that maximizes the log-probability, that is,

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

Log-likelihood function

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

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

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

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

which is the same as the loss used in the regression

- For N training samples $(\mathbf{x}^{(i)}, y^{(i)})$, by assuming they are *i.i.d.*, we can obtain their joint conditional pdf

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

- The log-likelihood function is

$$\log p(y^{(1)}, \dots, y^{(N)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) = -\frac{1}{2\sigma^2} \sum_{i=1}^N (y^{(i)} - \mathbf{x}^{(i)}\mathbf{w})^2 + \text{constant}$$

- Maximizing the log-likelihood $\log p(y^{(1)}, \dots, y^{(N)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$ is equivalent to minimize

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

which is the same as the loss used in the regression

- From the perspective of probability, linear regression is actually equivalent to
 - **Modeling:** assuming *conditional distribution to be 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

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

where $\mu \in [0, 1]$ is the probability of being 1

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

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

where $z = 0$ or 1

Binary Classification

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

$$p(y|x) = (\sigma(\mathbf{x}\mathbf{w}))^y \cdot (1 - \sigma(\mathbf{x}\mathbf{w}))^{1-y}$$

where $\mu = \sigma(\mathbf{x}\mathbf{w})$; and $y = 0$ or 1

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

$$\log p(y|x) = y \log \sigma(\mathbf{x}\mathbf{w}) + (1 - y) \log(1 - \sigma(\mathbf{x}\mathbf{w}))$$

Recall that the logistic regression minimizes

$$\text{cross entropy} \triangleq -y \log \sigma(\mathbf{x}\mathbf{w}) - (1 - y) \log(1 - \sigma(\mathbf{x}\mathbf{w}))$$

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

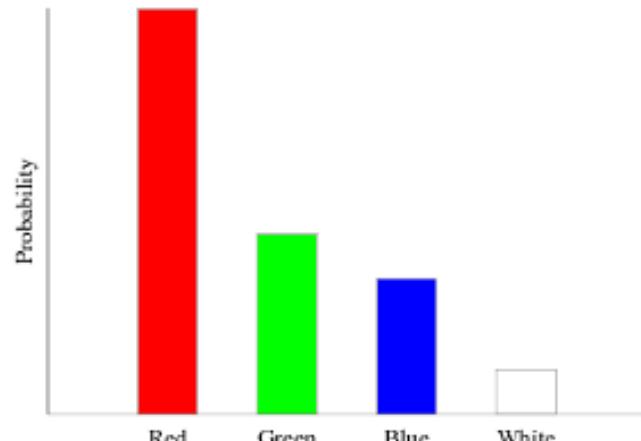
- The logistic regression is equivalent to
 - Modeling: assuming *Bernoulli conditional distribution* for the output
 - Training: training the model by *maximizing the log-likelihood*

Categorical Distribution

- The categorical distribution

$$p(\mathbf{z} = \text{onehot}_k) = \mu_k$$

- where $\text{onehot}_i = [0, \dots, 0, 1, 0, \dots, 0]$ is the a vector with the i -th element being the only nonzero element 1
- $\sum_{k=1}^K \mu_k = 1$



- The distribution can be equivalently written as

$$p(\mathbf{z}) = \prod_{k=1}^K \mu_k^{z_k}$$

where \mathbf{z} is a one-hot vector

Multiclass Classification

- **Modeling:** By setting the probability

$$\mu_k = \text{softmax}_k(\mathbf{x}\mathbf{W}),$$

the conditional probability distribution is assumed to be the categorical distribution

$$p(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^K [\text{softmax}_k(\mathbf{x}\mathbf{W})]^{y_k}$$

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

$$\log p(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^K y_k \cdot \log(\text{softmax}_k(\mathbf{x}\mathbf{W}))$$

$= - \text{cross entropy}$

Summary

- The regression, logistic and multi-class regressions are equivalent to
 - 1) assume different conditional pdfs for the outputs y
 - Regression: *Gaussian distribution*
 - Logistic regression: *Bernoulli distribution*
 - Multiclass logistic regression: *Categorical distribution*
 - 2) maximize the log-likelihood functions