# Statistical Learning for Data Science

## Lecture 14

唐晓颖
电子与电气工程系
南方科技大学

April 17, 2023

# Logistic Regression

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- Logistic regression for >2 response classes

Classify a response variable that has more than two classes (K classes).

$$\log\left(\frac{\Pr(Y=1\mid X)}{\Pr(Y=K\mid X)}\right) = \beta_{01} + \beta_{11}X_1 + \ldots + \beta_{p1}X_p$$

$$\log\left(\frac{\Pr(Y=2\mid X)}{\Pr(Y=K\mid X)}\right) = \beta_{02} + \beta_{12}X_1 + \ldots + \beta_{p2}X_p$$

.

.

$$\log\left(\frac{\Pr(Y=K-1\mid X)}{\Pr(Y=K\mid X)}\right) = \beta_{0K-1} + \beta_{1K-1}X_1 + \ldots + \beta_{pK-1}X_p$$

The model is specified in terms of K-1 log-odds or logits.

# Logistic Regression

- Logistic regression for >2 response classes

Classify a response variable that has more than two classes (K classes).

$$\Pr(Y=k \mid X)=\frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{0l}+\beta_{1l}X_1+\ldots+\beta_{pl}X_p}},\quad k=1,\ldots,K-1$$

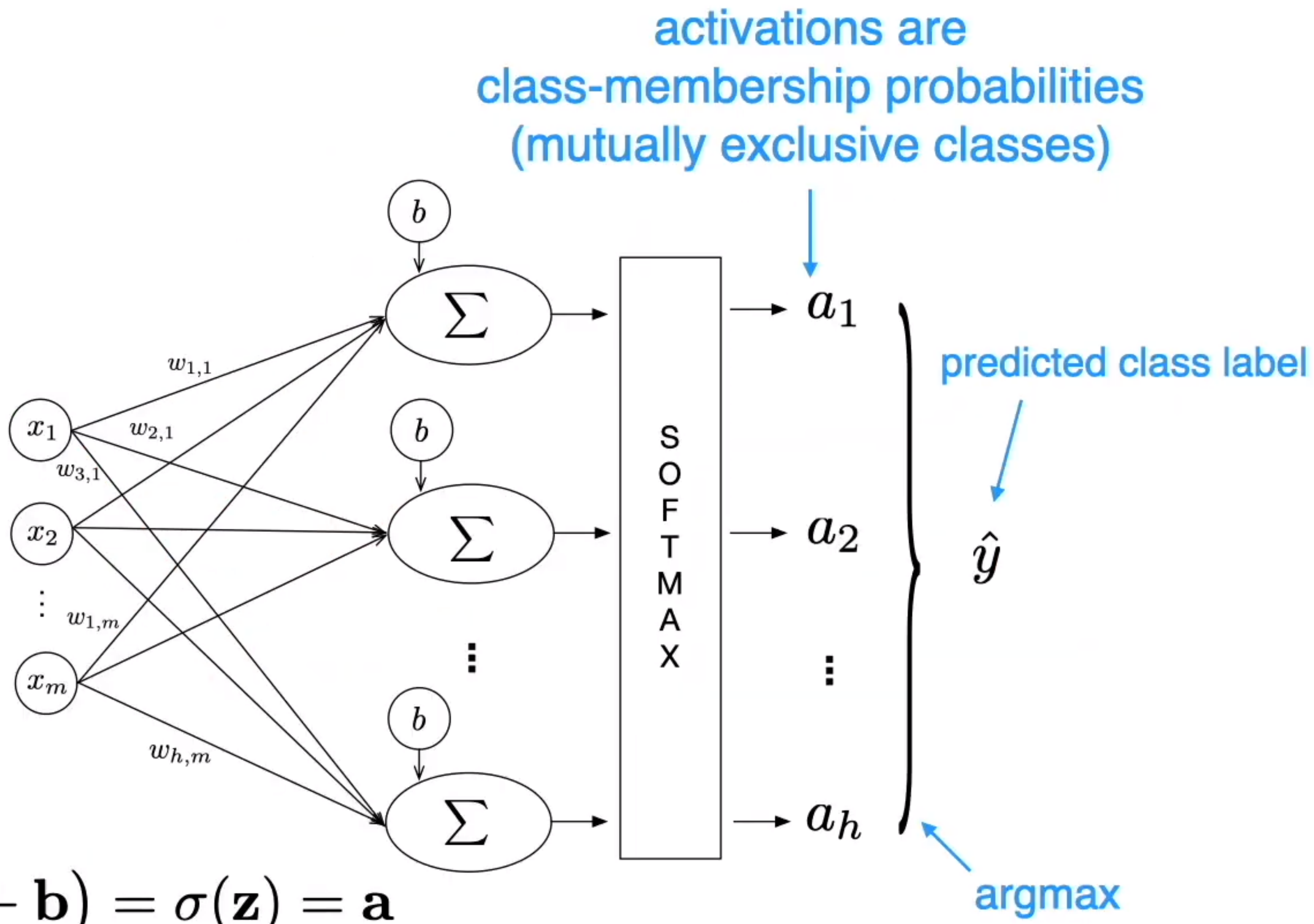
$$\Pr(Y=K \mid X)=\frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{0l}+\beta_{1l}X_1+\ldots+\beta_{pl}X_p}}$$

$$\sum_{l=1}^{K}\Pr(Y=l \mid X)=1$$

$$\hat{p}(X)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1X}}=\frac{e^{-10.6513+0.0055\times1000}}{1+e^{-10.6513+0.0055\times1000}}=0.00576$$

# Logistic Regression

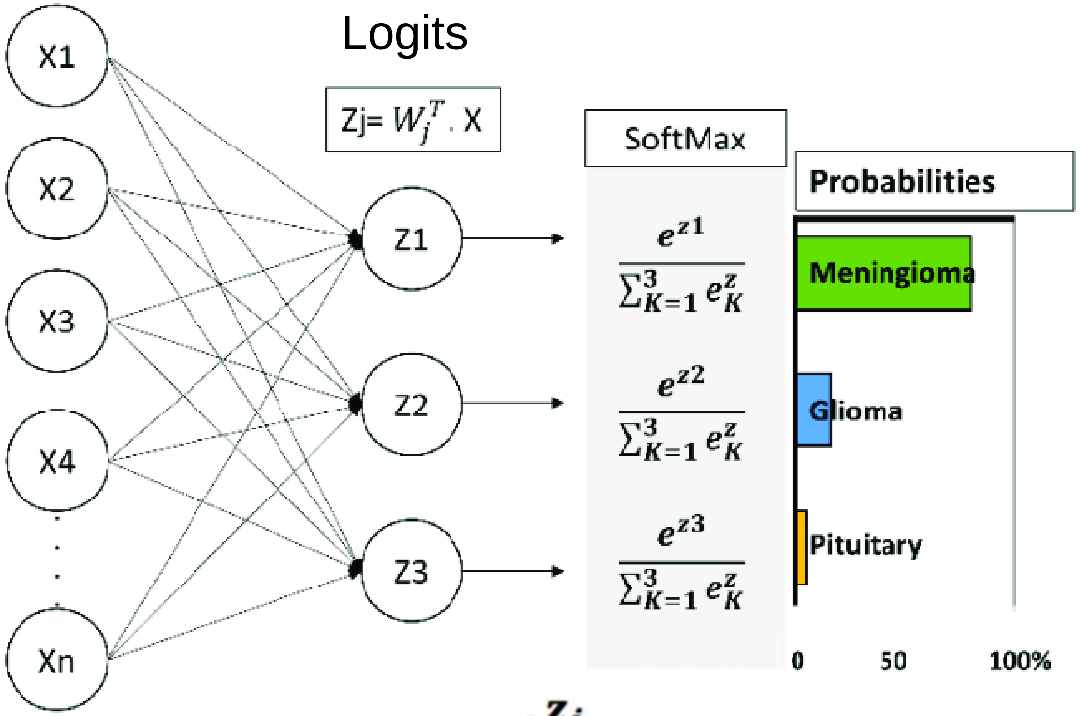
- Multinomial Logistic Regression/ SoftMax Regression




$$\sigma(\mathbf{Wx} + \mathbf{b}) = \sigma(\mathbf{z}) = \mathbf{a}$$

# Logistic Regression

- Relationship between Logistic Regression and SoftMax.




**SoftMax function:** $$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}} \; for\; j = 1, \ldots, k$$

The above is the softmax formula, which takes each Logits and find the probability. The numerator is the e-power values of the Logit and the denominator calculates the sum of the e-power values of all the Logits.

# Linear Discriminant Analysis (LDA)

In logistic regression, we model $\Pr(Y = k \mid X = x)$ using the logistic function (direct approach).

In discriminant analysis, we model the distribution of X in each of the classes separately $\Pr(X = x \mid Y = k)$, and then use *Bayes' theorem* to flip things around and obtain $\Pr(Y = k \mid X = x)$ (indirect approach)

# Linear Discriminant Analysis (LDA)

- Using Bayes' theorem for classification

$$\Pr(X = x \mid Y = k) \xleftrightarrow{\textit{Bayes' theorem}} \Pr(Y = k \mid X = x)$$

**HOW?**

$$\Pr(Y = k \mid X = x) = \frac{\Pr(X = x \mid Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

# Linear Discriminant Analysis (LDA)

$$\Pr(Y=k \mid X=x) = \frac{\Pr(X=x \mid Y=k) \cdot \Pr(Y=k)}{\Pr(X=x)}$$

- Using Bayes' theorem for classification

In discriminant analysis, we write

$$p_k(x) = \Pr(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

--- the *posterior* probability that an observation $X = x$ belongs to the *k*th class.

$f_k(x) = \Pr(X=x \mid Y=k)$ --- the *density* for $X$ in class $k$. In LDA, we will use normal densities, for these, separately in each class.

$\pi_k = \Pr(Y=k)$ --- the marginal or *prior* probability for $k$ class.

- A simple estimate of the **prior**: the fraction of the training observations that belong to the *k*-th class.
- Estimating the **density** is more challenging, unless we assume some simple forms for these densities.

贝叶斯方法的核心在于： **后验概率 = 先验概率 * 来自数据的信息**。

例子：假设A和B和C是3个人，3个人都不认识；让A和B打牌 ，C来猜谁赢的可能性大；没打牌之前C会猜A和B赢的几率各为50%；刚打完一次 A赢了，这时候C就会认为A的技术可能会更好赢的几率会大于50%，A和B继续打牌 一会A连续几次赢一会B连续赢几次，C在这个过程中有时候认为A技术好 点有时候会认为B技术好点；C的判断随着打牌次数不断变化，就是贝叶斯概率原理。

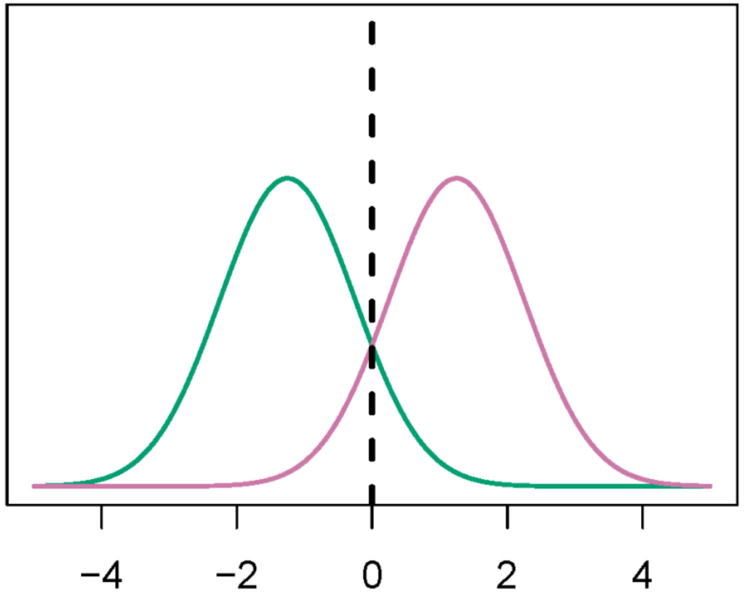
贝叶斯方法最好的地方就在于：你不用知道全局，甚至你一开始可以错的离谱，但你可以越来越接近真理。

# Linear Discriminant Analysis (LDA)

- Using Bayes' theorem for classification

$$p_k(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

$\pi_1$=.5, $\pi_2$=.5

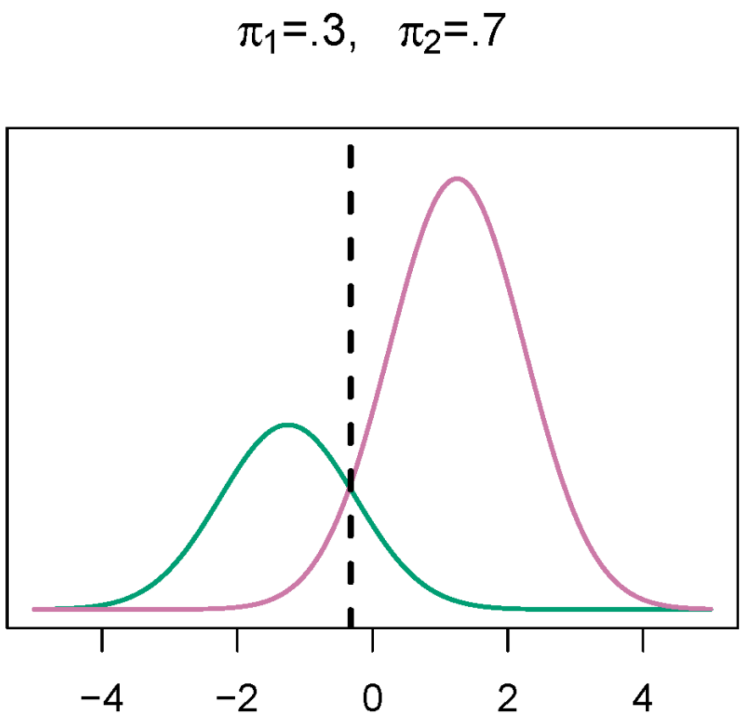



When the priors are equal, we classify a new point according to which density is highest.

# Linear Discriminant Analysis (LDA)

- Using Bayes' theorem for classification

$$p_k(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$




When the priors are different, we take them into account as well, and compare $\pi_k f_k(x)$.

# Linear Discriminant Analysis (LDA)

- Using Bayes' theorem for classification

$$p_k(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Comment: The Bayes classifier classifies an observation to the class for which $p_k(x)$ is largest. It has the lowest possible error rate out of all classifiers. Therefore, if we can find a way to estimate $f_k(X)$ , then we can develop a classifier that approximates the Bayes classifier.

# Linear Discriminant Analysis (LDA)

- Why discriminant analysis, given that we already have logistic regression?

1. When the classes are well-separated, the parameter estimates for the logistic regression model are surprising unstable. LDA does not suffer from this problem.

2. If n is small and the distribution of X is approximately normal in each of the classes, LDA is again more stable than logistic regression.

3. LDA is popular when we have more than two response classes.

# Linear Discriminant Analysis (LDA)

- The main idea of LDA:

1. In LDA, we assume that the data is normally distributed and that each class has its own mean and covariance matrix.

2. The goal of LDA is to find a projection of the data that maximizes the separation between the classes while minimizing the variance within each class.

3. LDA can also be used for dimensionality reduction such as reducing the number of features in a dataset.

# Linear Discriminant Analysis (LDA)

- LDA for p=1

When p=1, the Gaussian (normal) density has the form

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left( -\frac{1}{2\sigma_k^2}(x-\mu_k)^2 \right)$$

$\mu_k$ and $\sigma_k^2$ are the mean and variance parameters for the *k*th class. We will assume that all the $\sigma_k^2 = \sigma^2$ are the same.

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{1}{2\sigma^2}(x-\mu_k)^2 \right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{1}{2\sigma^2}(x-\mu_l)^2 \right)}$$

# Linear Discriminant Analysis (LDA)

- LDA for p=1

To classify at the value of $X = x$ , we need to see which of the $p_k(x)$ is largest. Taking logs, and discarding terms that do not depend on $k$, we see that this is equivalent to assigning $x$ to the class with the largest *discriminant score*:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^{\ 2}}{2\sigma^2} + \log(\pi_k)$$

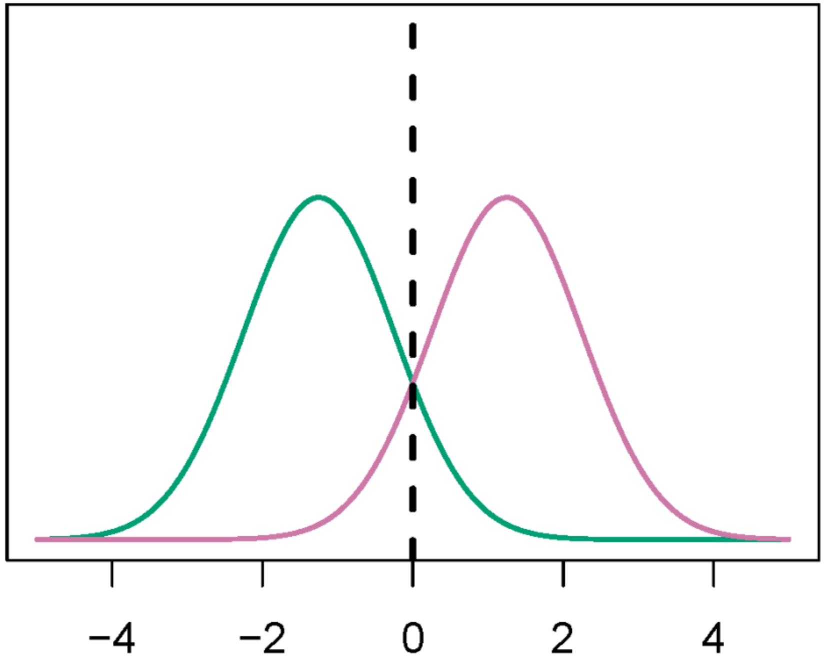
$\delta_k(x)$ is a *linear* function of $x$ --- that's why it is called **linear** DA!

Question: If there are two classes with equal prior probabilities, then what is the decision boundary?

$$x = \frac{\mu_1 + \mu_2}{2}$$

# Linear Discriminant Analysis (LDA)

- LDA for p=1




$$\mu_1 = -1.25, \mu_2 = 1.25$$
$$\sigma_1^2 = \sigma_2^2 = 1$$
$$\pi_1 = \pi_2 = 0.5$$

- The Bayes classifier assigns the observation to class 1 if $x<0$ and class 2 otherwise.
- In this case, we can compute the Bayes classifier because we know that X is drawn from a Gaussian distribution within each class, and we know all the parameters involved. In real-life situation, we are not able to calculate the Bayes classifier.
- In practice, under the assumption of normal distributions, we still need to estimate the parameters $\mu_1, \mu_2, \ldots, \mu_K$, $\pi_1, \pi_2, \ldots, \pi_K$ and $\sigma^2$.