UVA CS 4501: Machine Learning

Lecture 13: Logistic Regression

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Where are we ? ➔ Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

Where are we ? → Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., logistic regression, support vector machine, decisionTree

2. Generative:
   - build a generative statistical model
   - e.g., naïve bayes classifier, Bayesian networks

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

Today

- Bayes Classifier
- Logistic Regression
- Binary to multi-class
- Training LG by MLE

Bayes classifiers

- Treat each feature attribute and the class label as random variables.

Bayes classifiers

- Treat each feature attribute and the class label as random variables.

- Given a sample **x** with attributes ($x_1, x_2, \ldots, x_p$):
  - Goal is to predict its class *c.*
  - Specifically, we want to find the class that maximizes $p(c \mid x_1, x_2, \ldots, x_p)$.

MAP Rule

Bayes Classifiers – MAP Rule

Task: Classify a new instance X based on a tuple of attribute values $X = \langle X_1, X_2, \ldots, X_p \rangle$ into one of the classes

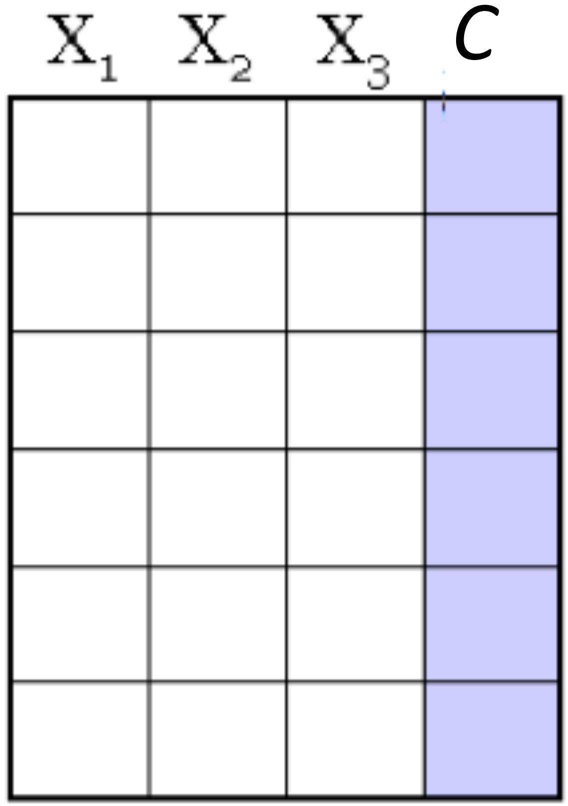
$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid x_1, x_2, \ldots, x_p)$$

MAP Rule

MAP = Maximum Aposteriori Probability

Please read the L13Extra slides for WHY

Adapt From Carols' prob tutorial



A Dataset for classification

$$f : \boxed{X} \longrightarrow \boxed{C}$$

Output as Discrete Class Label
$C_1, C_2, \ldots, C_L$

Discriminative

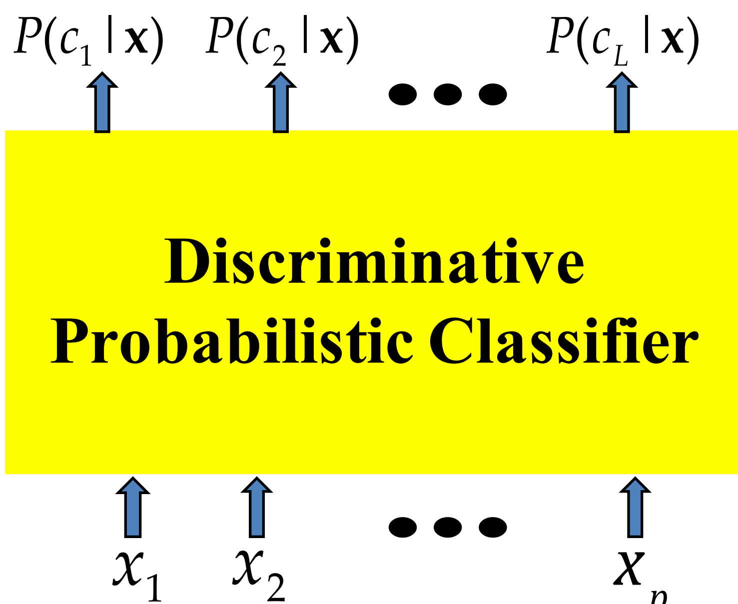
$$\underset{c \in C}{\arg\max} P(c / \mathbf{X}) \quad C = \{c_1, \cdots, c_L\}$$

- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

Establishing a probabilistic model for classification

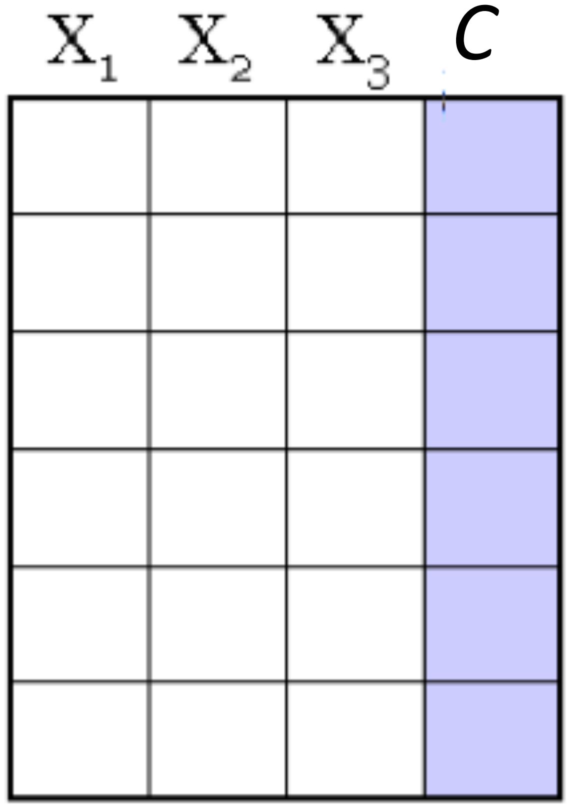
– **Discriminative model**

$$\arg\max_{c \in C} P(c/\mathbf{X}), \quad C = \{c_1, \cdots, c_L\}$$



$$\mathbf{x} = (x_1, x_2, \cdots, x_p)$$

Adapt from Prof. Ke Chen NB slides



A Dataset for classification

$$f : X \longrightarrow C$$

Output as Discrete Class Label
$C_1, C_2, \ldots, C_L$

Discriminative

$$\underset{c \in C}{\arg\max}\, P(c / \mathbf{X}) \quad C = \{c_1, \cdots, c_L\}$$

Generative

$$\underset{c \in C}{\operatorname{argmax}} P(c|X) = \underset{c \in C}{\operatorname{argmax}} P(X, c) = \underset{c \in C}{\operatorname{argmax}} P(X|c)P(c)$$

Later!

- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

Today

- ❑ Bayes Classifier
- ❑ Logistic Regression
- ❑ Binary to multi-class
- ❑ Training LG by MLE

Multivariate linear regression to Logistic Regression

$$\mathbf{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

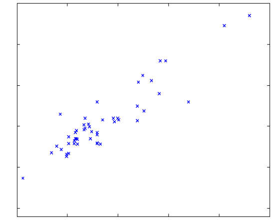
linear

Logistic regression for binary classification

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

logistic

Review: Probabilistic Interpretation of Linear Regression



- Let us assume that the target variable and the inputs are related by the equation:

$$y_i = \theta^T \mathbf{x}_i + \varepsilon_i$$

RV $\varepsilon \sim N(0, \sigma^2)$

where ε is an error term of unmodeled effects or random noise

- Now assume that ε follows a Gaussian $N(0,\sigma^2)$, then we have:

$$p(y_i \mid x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T \mathbf{x}_i)^2}{2\sigma^2} \right)$$

RV $y \mid x; \theta \sim N(\theta^T x, \sigma)$

Logistic Regression p(y|x)

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right]=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p$$

$$\Downarrow$$

$$\mathrm{P}(\mathrm{y}|\mathrm{x})=\frac{\mathrm{e}^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p}}{1+\mathrm{e}^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p}}=\frac{1}{1+\mathrm{e}^{-\beta^TX}}$$

Logistic Regression models a linear classification boundary!

$$y \in \{0,1\}$$

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Logistic Regression models a linear classification boundary!

[Seperate two classes]

$$\ln \frac{P(y=1|x)}{1-P(y=1|x)} = \ln \frac{P(y=1|x)}{P(y=0|x)} = 0$$

0.5

0.5

linear hyperplane

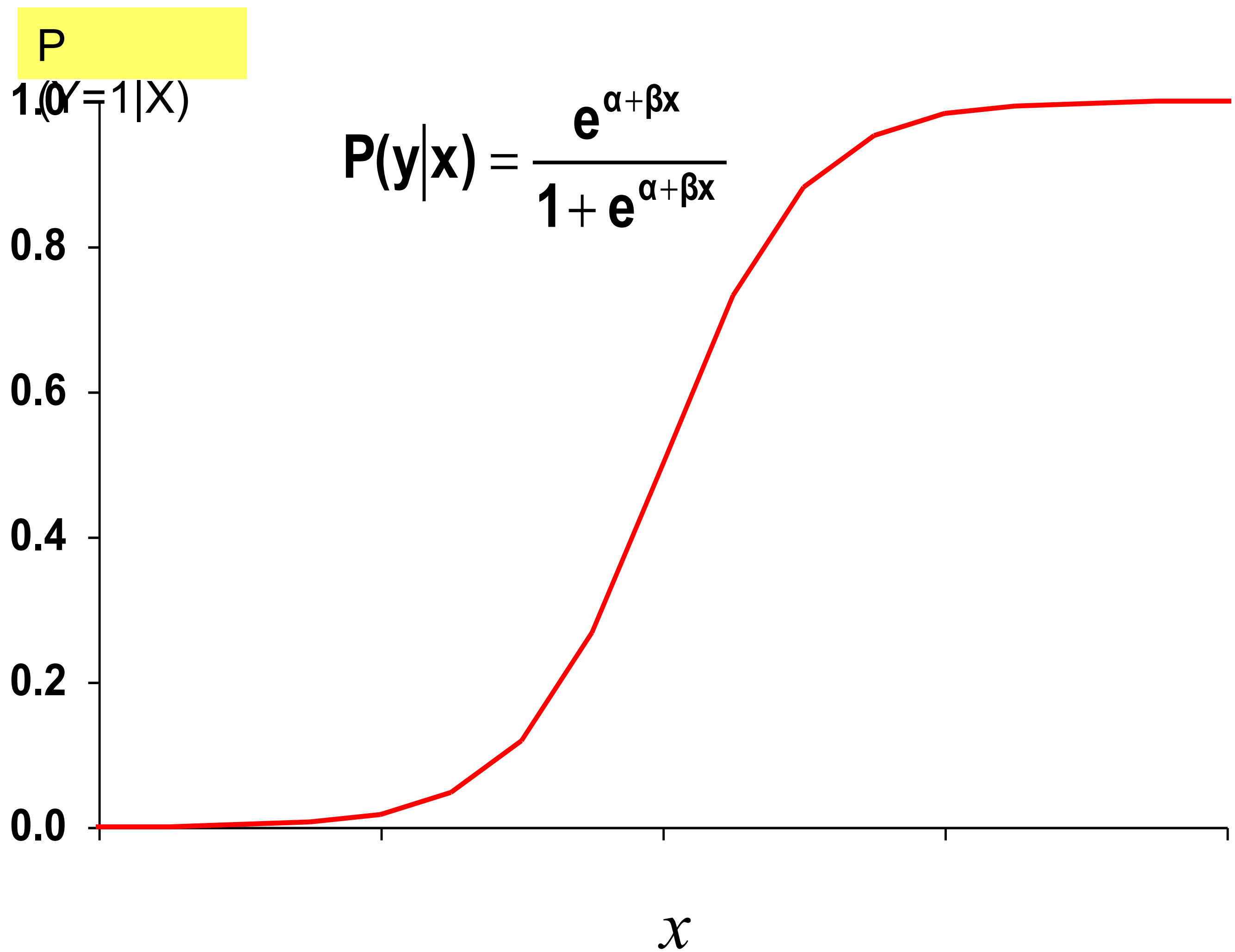
$$\alpha + \beta_1 x_1 + \dots + \beta_p x_p = 0$$

Boundary points

$$p(y=1|x) = p(y=0|x)$$

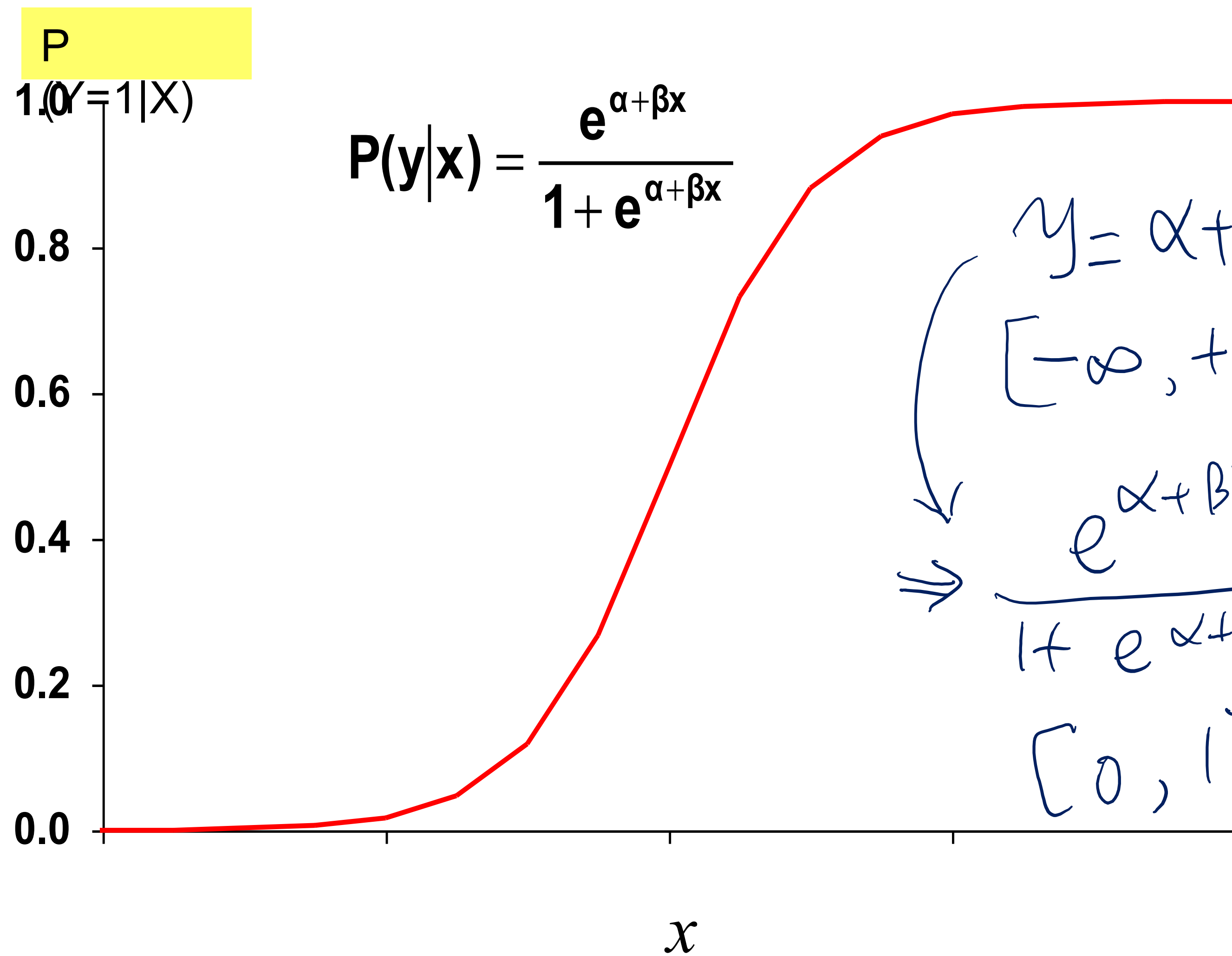
The logistic function (1) -- is a common "S" shape function

e.g. Probability of disease



The logistic function (1) -- is a common "S" shape function

e.g. Probability of disease



$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$y = \alpha + \beta x$

$[-\infty, +\infty]$

$\Rightarrow \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$

$[0, 1]$

Logistic Regression—when?

Logistic regression models are appropriate for target variable coded as 0/1.

We only observe "0" and "1" for the target variable—but we think of the target variable conceptually as a probability that "1" will occur.

P(x) 1-p(x)

19

Logistic Regression—when?

Logistic regression models are appropriate for target variable coded as 0/1.

⇒ y is model with Bernoulli (p)

We only observe "0" and "1" for the target variable—but we think of the target variable conceptually as a probability that "1" will occur.

⇒ p is a func of x

This means we use Bernoulli distribution to model the target variable with its Bernoulli parameter p=p(y=1|x) predefined.

The main interest ➔ predicting the probability that an event occurs (i.e., the probability that p(y=1|x)).

The logit function View

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

logistic ʃ

$$\underbrace{\ln\left[\frac{P(y|x)}{1-P(y|x)}\right]}_{\text{Logit of } P(y|x)} = \alpha + \beta x$$

logit / log-odd

Logit function

Decision Boundary ➔ equals to zero

$$\ln\left[\frac{P(y=1|x)}{P(y=0|x)}\right] = \ln\left[\frac{P(y=1|x)}{1-P(y=1|x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Logistic Regression Assumptions

- Linearity in the logit – the regression equation should have a linear relationship with the logit form of the target variable

- There is no assumption about the feature variables / target predictors being linearly related to each other.

Today

- ❑ Bayes Classifier
- ❑ Logistic Regression
- ❑ Binary to multi-class
- ❑ Training LG by MLE

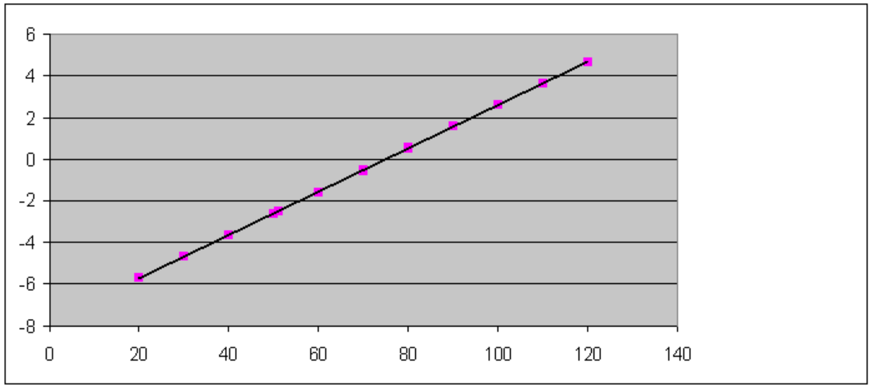
Binary Logistic Regression (K=2)

In summary that the logistic regression tells us two things at once.

- Transformed, the "log odds" (logit) are linear.

ln[p/(1-p)]

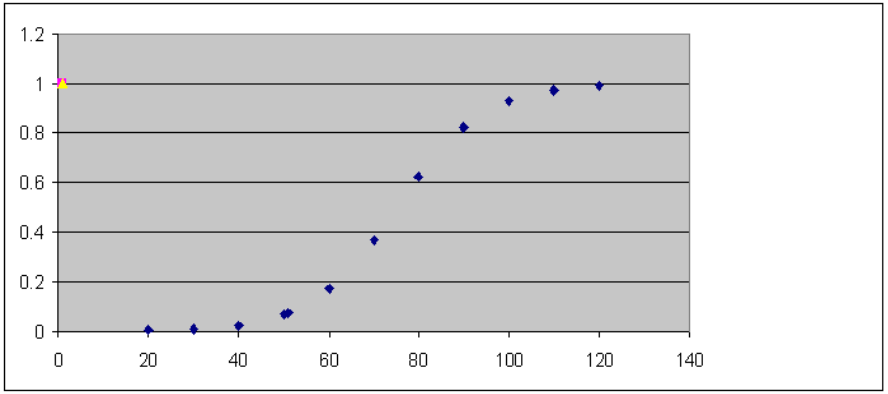
Odds= p/(1-p)



x

- Logistic Distribution

P (Y=1|x)

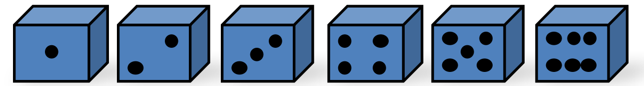


x

This means we use Bernoulli distribution to model the target variable with its Bernoulli parameter p=p(y=1|x) predefined.

P(y=1|x) 1-p(x)

Binary → Multinoulli Logistic Regression Model



$p(G|x)$

Directly models the posterior probabilities as the output of regression

$$\Pr(G=k \mid X=x) = \frac{\exp(\beta_{k0}+\beta_k^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)},\ k=1,\ldots,K-1$$

$$\Pr(G=K \mid X=x) = \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)}$$

x is p-dimensional input vector

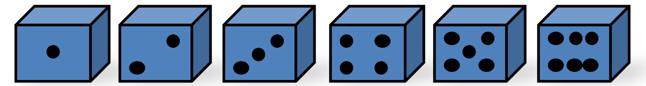
β_k^T is a p-dimensional vector for each class k

Total number of parameters is $(K\text{-}1)(p\text{+}1)$ $\beta_{k0}, \vec{\beta}_k$, $k=1,2,\ldots,K-1$

Note that the class boundaries are linear

Binary → Multinoulli Logistic Regression Model

(e.g. $K=6$)



$p(y=1|x) \Rightarrow \frac{e^{\beta x}}{1+e^{\beta x}}$

$p(y=0|x) \Rightarrow \frac{1}{1+e^{\beta x}}$

$$\frac{e^{\beta_k^T x}}{1+e^{\beta_1^T x}+e^{\beta_2^T x}\cdots}$$

e.g.

$$\ln \frac{P(G=k|x)}{P(G=K|x)} = 0 \Rightarrow \text{linear}$$

$$\ln \frac{P(G=k|x)}{P(G=K|x)} = \beta_{k0}+\beta_k^T x$$

Note that the class boundaries are linear

Today

- ❑ Bayes Classifier
- ❑ Logistic Regression
- ❑ Binary to multi-class
- ❑ Training LG by MLE

Parameter Estimation for LG → MLE from the data

- Review:
  - Linear regression training → Least squares
  - Gaussian explaination of Linear Regression training → MLE
- Logistic regression Training:
  - Maximum Likelihood Estimation based
  - Classic algorithms using iterative Newton

Please read the L13Extra slides for HOW

Review: Defining Likelihood for basic Bernoulli

- Likelihood = p(data | parameter)

➔ e.g., for n independent tosses of coins, with unknown p

function of x_i

PMF: $$f(x_i \mid p) = p^{x_i}(1-p)^{1-x_i}$$

$$x = \sum_{i=1}^{n} x_i$$

LIKELIHOOD:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{x}(1-p)^{n-x}$$

function of p

Observed data ➔ x heads-up from n trials

MLE for Logistic Regression Training (extra)

Let's fit the logistic regression model for K=2, i.e., number of classes is 2

Training set: (x_i, y_i), i=1,…,N

For Bernoulli distribution

$$p_\beta(y|x)^y(1-p_\beta(y|x))^{1-y}$$

How?

$$l(\beta)=\sum_{i=1}^{N}\{\log\Pr(Y=y_i|X=x_i)\}$$

$$=\sum_{i=1}^{N}y_i\log(\Pr(Y=1|X=x_i))+(1-y_i)\log(\Pr(Y=0|X=x_i))$$

$$=\sum_{i=1}^{N}(y_i\log\frac{\exp(\beta^Tx_i)}{1+\exp(\beta^Tx_i)})+(1-y_i)\log\frac{1}{1+\exp(\beta^Tx_i)})$$

$$=\sum_{i=1}^{N}(y_i\beta^Tx_i-\log(1+\exp(\beta^Tx_i)))$$

x_i are (p+1)-dimensional input vector with leading entry 1
\beta is a (p+1)-dimensional vector

We want to maximize the log-likelihood in order to estimate *beta*

y_i

$p(y_i = 1 | x)$

$$l(\beta) = \sum_{i=1}^{N} \{\log \Pr(Y = y_i | X = x_i)\}$$

$$\log\{\Pr(Y = y_i | X = x_i) = p(y_i | x_i)\} \Rightarrow \begin{cases} y_i = 1 \\ y_i = 0 \end{cases}$$

$1 - p(y_i = 1 | x)$

$$= \log\{p(y_i = 1 | x)^{y_i} (1 - p(y_i = 1 | x_i))^{1-y_i}\}$$

$$= y_i \log p(y_i = 1 | x) + (1 - y_i) \log(1 - p(y_i = 1 | x))$$

Logistic Regression

Task → Representation → Score Function → Search/Optimization → Models, Parameters

classification → Log-odds = linear function of X' s → EPE, MLE → Iterative (Newton) method → Logistic weights

$$P(c=1|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

References

- Prof. Tan, Steinbach, Kumar's "Introduction to Data Mining" slide
- Prof. Andrew Moore's slides
- Prof. Eric Xing's slides
- Prof. Ke Chen NB slides
- Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.