Machine Learning

Lecture 14: Logistic Regression

Dr. Beilun Wang

Southeast University
School of Computer Science
and Engineering

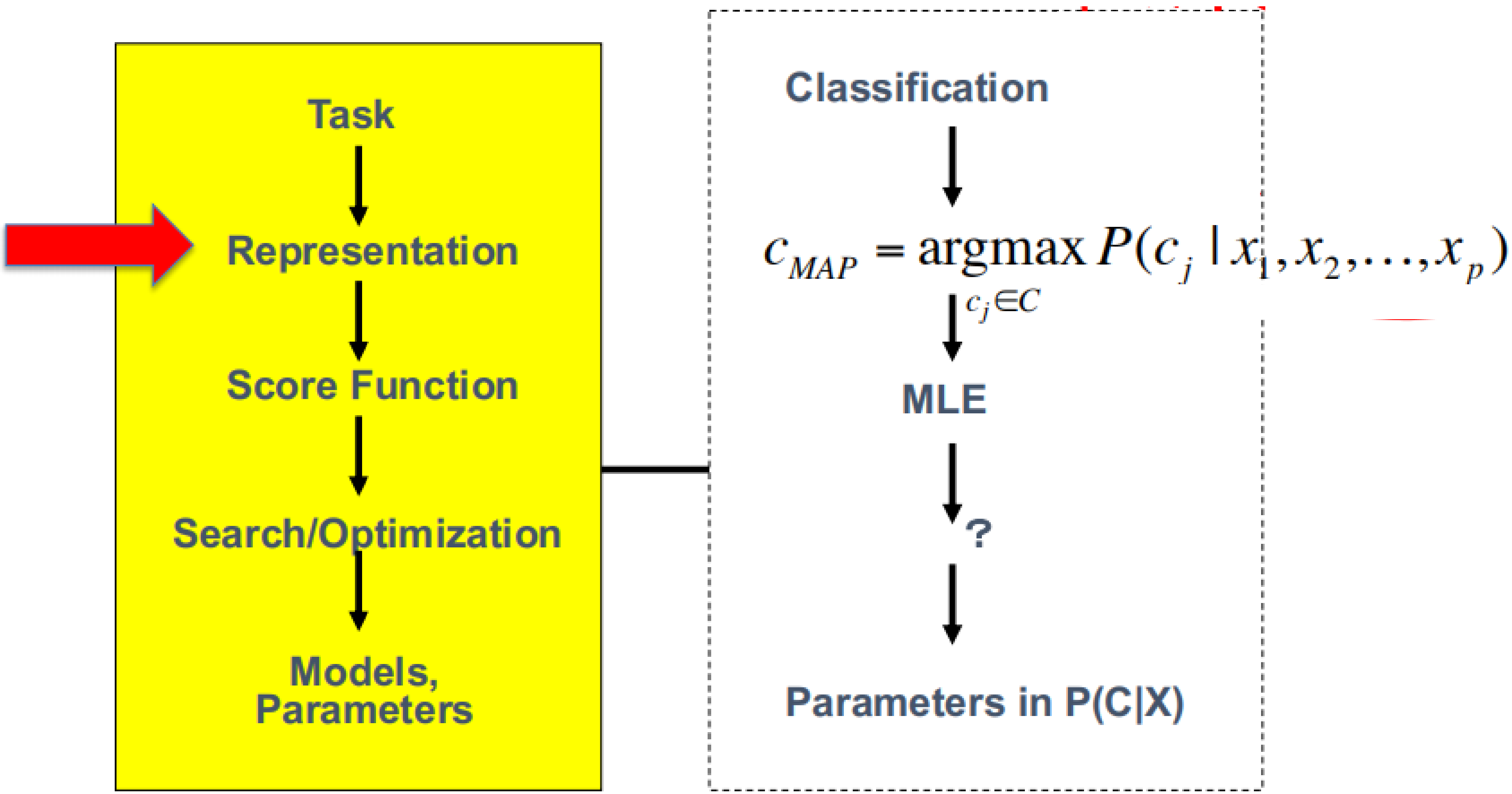
Course Content Plan

- ❑ Regression (supervised) — Y is a continuous
- ❑ Classification (supervised) — Y is a discrete
- ❑ Unsupervised models — NO Y
- ❑ Learning theory — About f()
- ❑ Graphical models — About interactions among X1,… Xp
- ❑ Reinforcement Learning — Learn program to Interact with its environment

Today

- Bayes Classifier
- Logistic Regression
- Training LG by MLE

Bayes Classifier



Bayes Classifiers

- Treat each feature attribute and the class label as random variables.
- Testing: Given a sample $\boldsymbol{x}$ 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 $P(c|x_1, x_2, \ldots, x_p)$.
- Training: can we estimate

$$P(c|\boldsymbol{x}) = P(c|x_1, x_2, \ldots, x_p)$$

directly from data?

Bayes Classifiers – MAP Rule

- *Task*: Classify a new instance X based on a tuple of attribute values $X = (X_1, X_2, \dots, X_p)$ into one of the classes

$$c_{MAP} = argmax_{c_j \in C} P(c_j | x_1, x_2, \dots, x_p)$$

- MAP = Maximum A posteriori Probability

Bayes Classifiers – MAP Classification Rule

- Establishing a probabilistic model for classification
  → MAP classification rule
  - **MAP**: **M**aximum **A** **P**osterior
  - Assign ***x*** to *c** if

$$P(C = c^* | \mathbf{X} = \mathbf{x}) > P(C = c | \mathbf{X} = \mathbf{x})$$
$$\text{for } c \neq c^*, \; c = c_1, \cdots, c_L$$

Review: Statistical Decision Theory

- Random input vector: X
- Random output variable: Y
- Joint distribution: Pr(X,Y)
- Loss function L(Y, f(X))
- Expected prediction error (EPE):

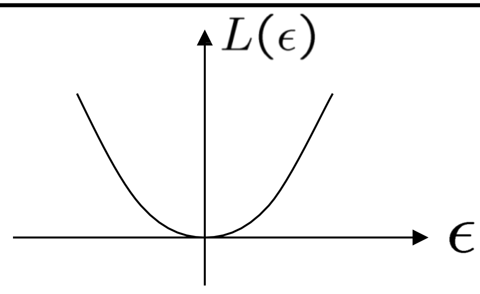
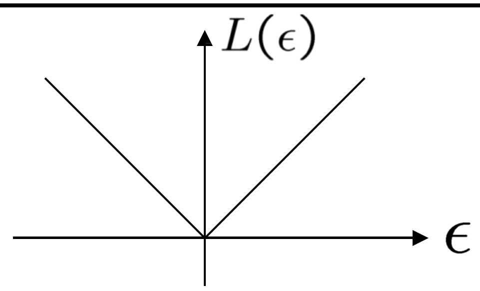
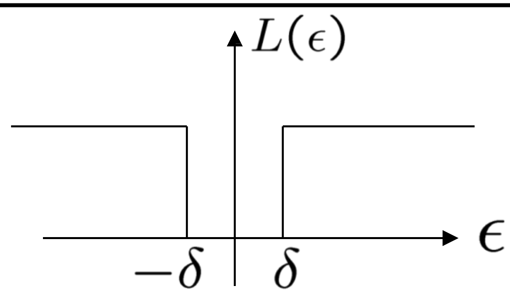
$$\mathrm{EPE}(f) = \mathrm{E}(L(Y, f(X))) = \int L(y, f(x))\Pr(dx, dy)$$

$$\text{e.g.} = \int (y - f(x))^2 \Pr(dx, dy)$$

e.g. Squared error loss (also called L2 loss)

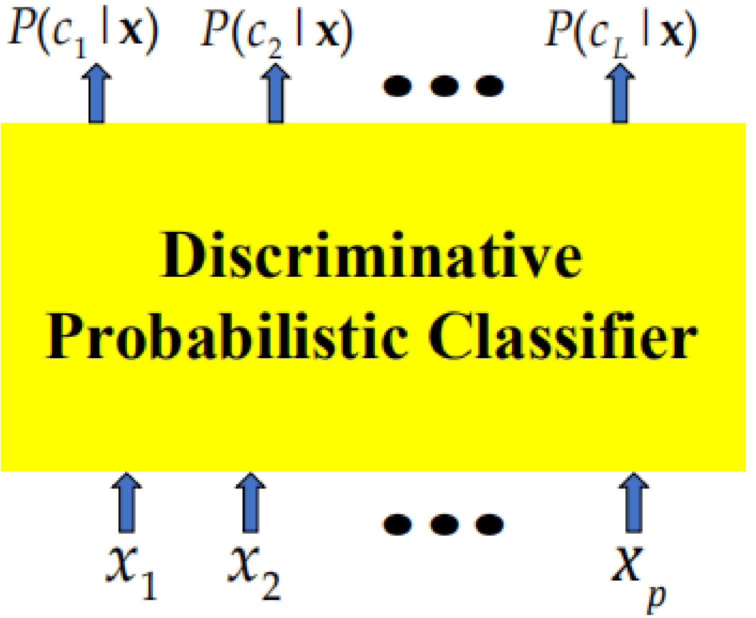
Consider population distribution

Review: EPE with different loss

Loss Function		Estimator $\hat{f}(x)$
L_2		$\hat{f}(x) = E[Y \mid X = x]$
L_1		$\hat{f}(x) = \text{median}(Y \mid X = x)$
0-1		$\hat{f}(x) = \arg\max_{Y} P(Y \mid X = x)$ (Bayes classifier / MAP)

Today: Discriminative model

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

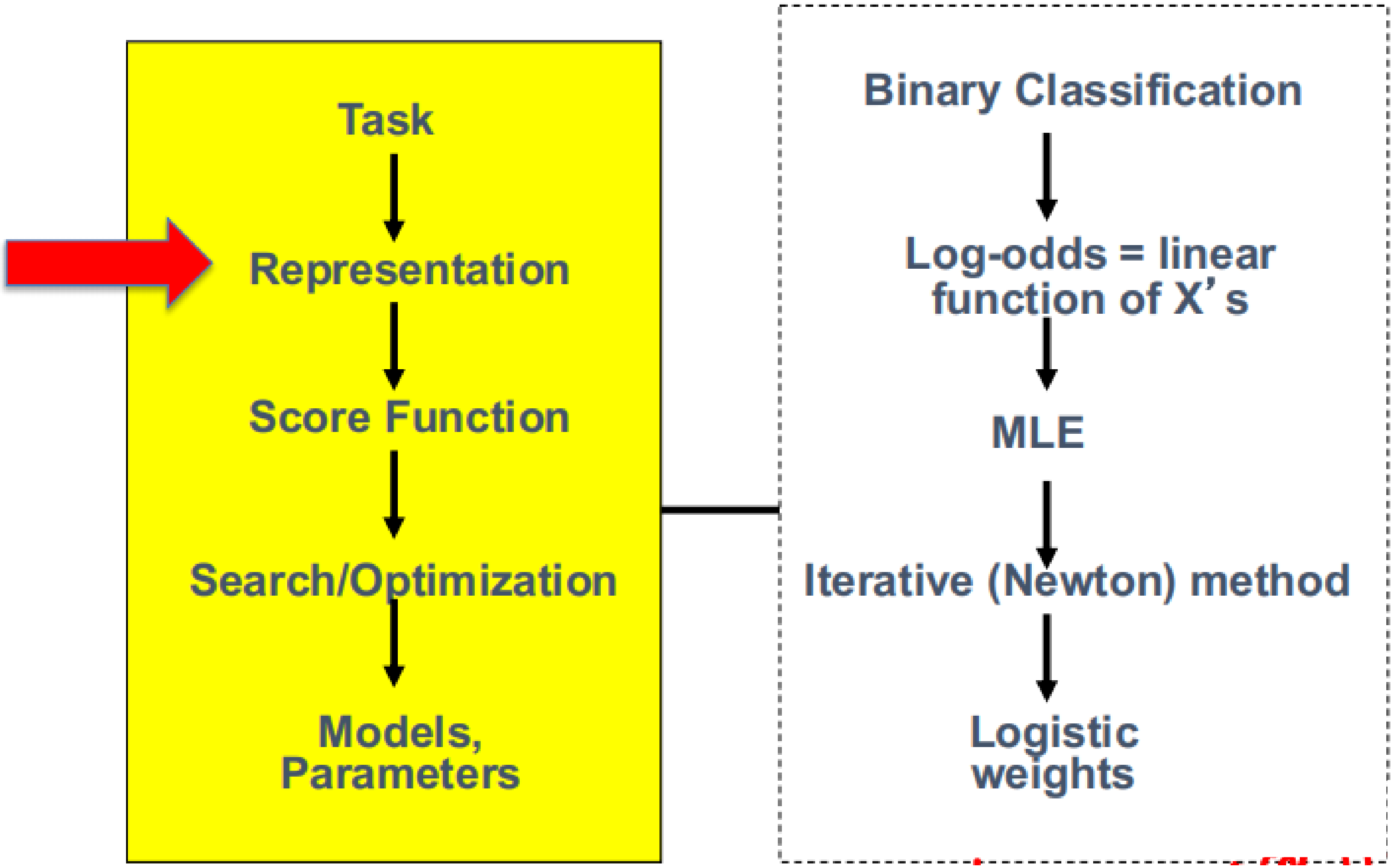


$$\mathbf{x} = (x_1, x_2, \cdots, x_p)$$

Today

- Bayes Classifier
- Logistic Regression
- Training LG by MLE

Logistic Regression



Multivariate linear regression to Logistic Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Logistic regression for binary classification

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

Logistic Regression $\mathrm{P}(y|x)$

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

$$\Downarrow$$

$$P(y|x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} = \frac{1}{1+e^{-(\beta_0+\beta^T X)}}$$

The logit function View (e.g. when with 1D x)

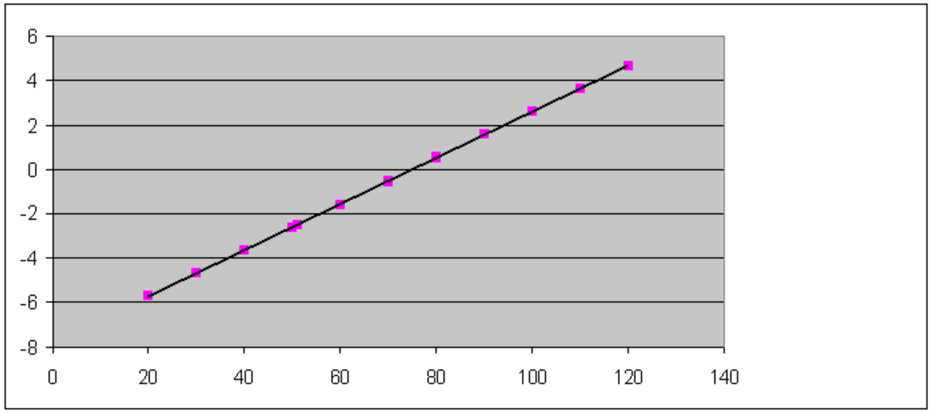
$$P(y|x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

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

Logit function

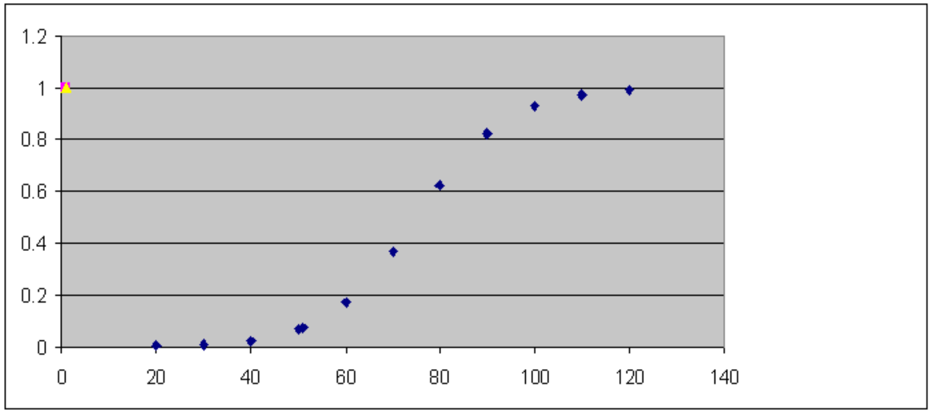
Binary Logistic Regression: Three Views

$$\ln[\frac{P}{1-P}]$$



x

$$P(y=1|x)$$



x

Bernoulli distribution

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

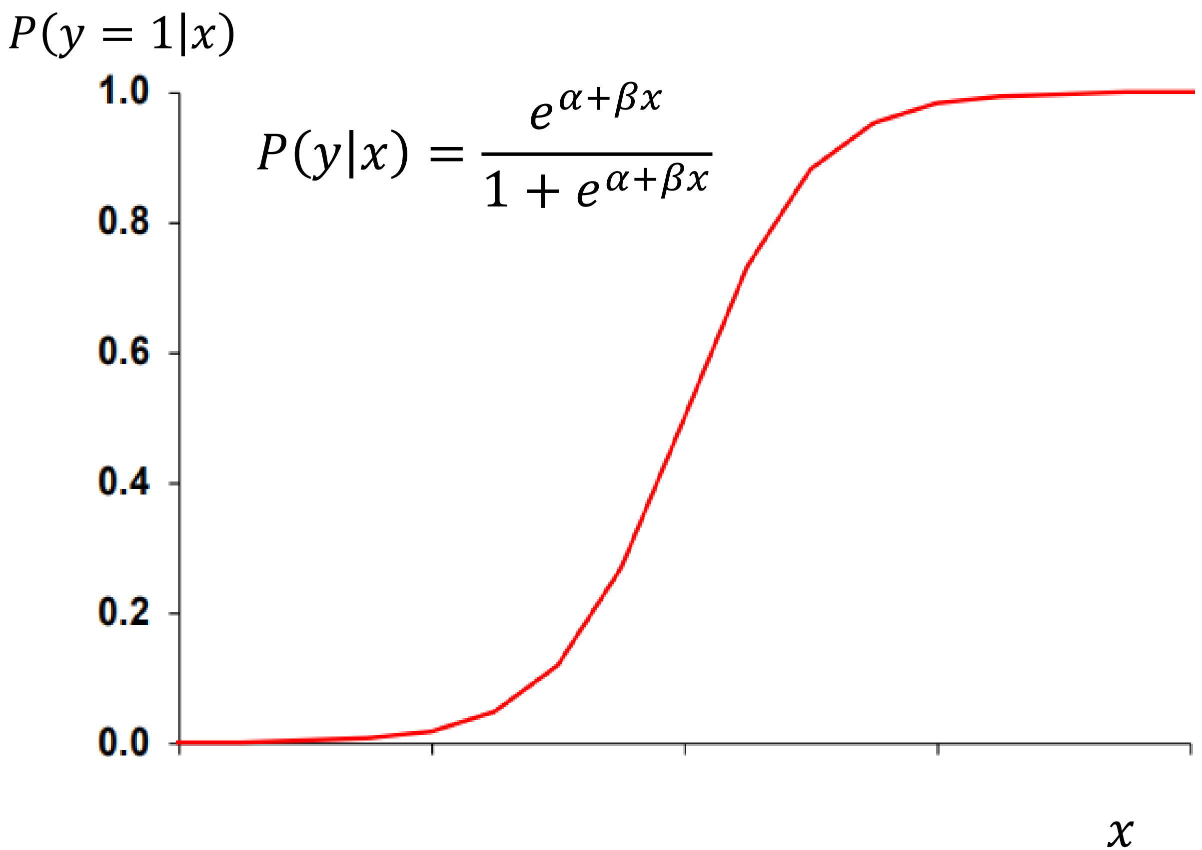
$P(y=1|x)$ $\quad$ $1-P(y=1|x)$

$$P_{Head} =$$

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

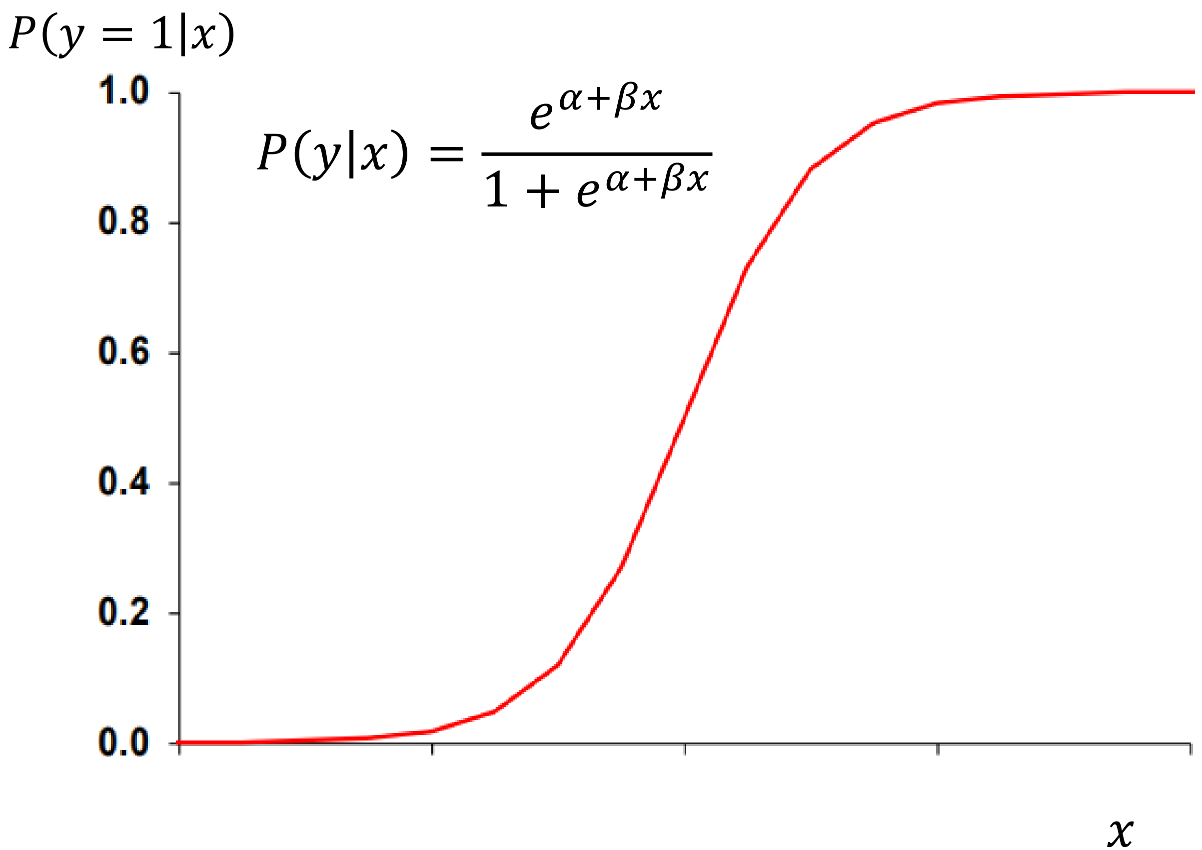
Binary Logistic Regression (Three Views)

- View I: logit of $P(y = 1|x)$ is linear function of x



Binary Logistic Regression (Three Views)

- View II: "S" shape function compress output to [0,1]



Binary Logistic Regression (Three Views)

- View III: Logistic Regression models a linear classification boundary

$$y \in \{0, 1\}$$
$$argmax_{y\in\{0,1\}} P(y|x)$$

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

$$\log\left[\frac{P(y = 1|x)}{P(y = 0|x)}\right] = \vec{\beta}^T\vec{x} = \log(1) = 0$$

Binary Logistic Regression (Three Views)

- View III: Logistic Regression models a linear classification boundary

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

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

$$\ln\left[\frac{P(y=1|x)}{P(y=0|x)}\right] = \ln\left[\frac{P(y=1|x)}{1 - P(y=1|x)}\right]$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Decision Boundary ➔ equals to zero

When to use Logistic Regression?

- Logistic regression models are appropriate when the target variable is 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.

Logistic Regression Assumptions

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

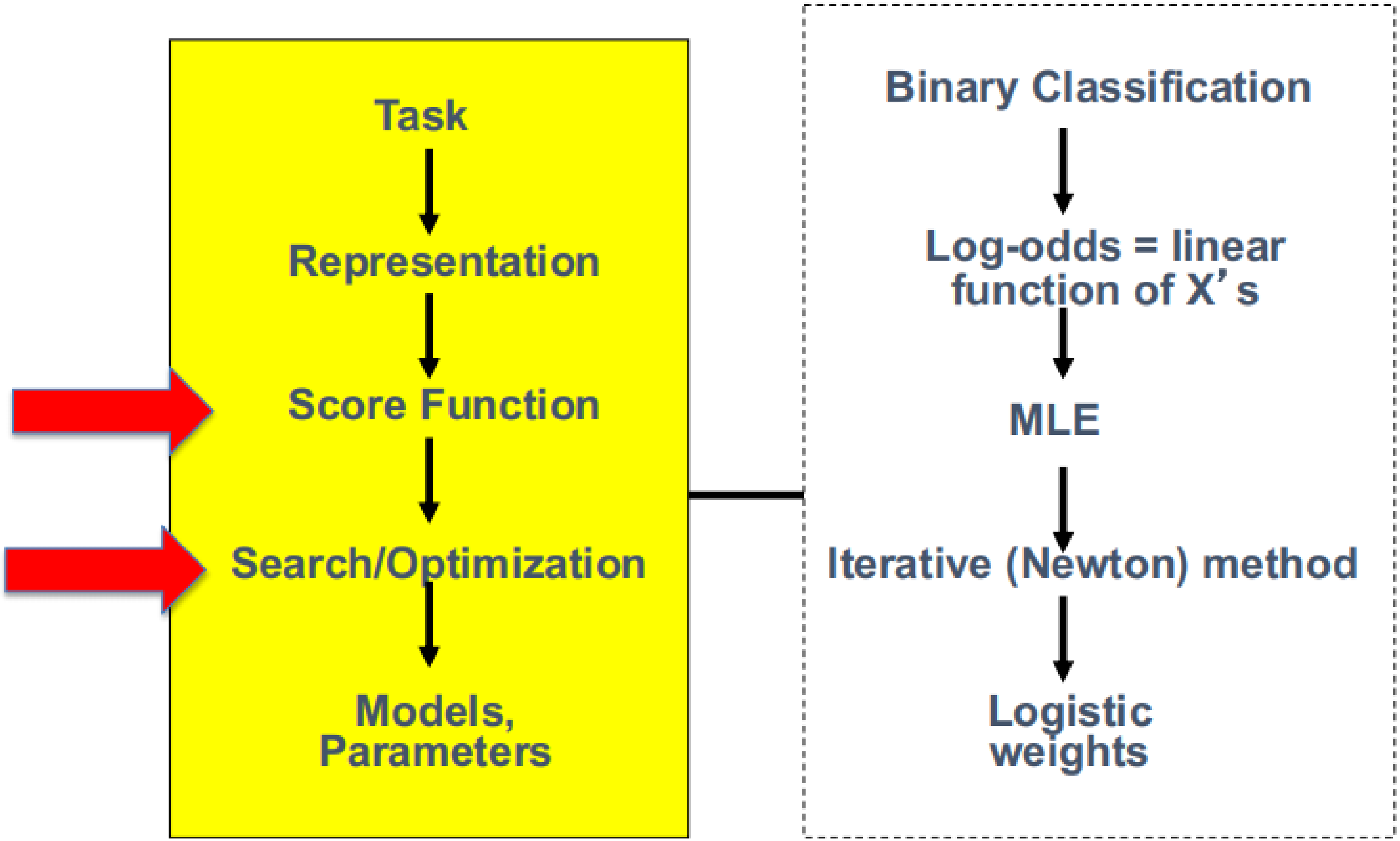
- There is no assumption about the feature variables / target predictors being linearly related to each other.

$P(y = 1|x)$ $\quad$ $1 - P(y = 1|x)$

Today

- Bayes Classifier
- Logistic Regression
- Training LG by MLE

Logistic Regression



Review: Maximum Likelihood Estimation

- Consider a sample set $T = (Z_1, \dots, Z_n)$ which is drawn from a probability distribution $P(Z|\theta)$ where θ are parameters.
- If the Zs are independent with probability density function $P(Z_i|\theta)$, the joint probability of the whole set is

$$P(Z_1, \dots, Z_n|\theta) = \prod_{i=1}^{n} P(Z_i|\theta)$$

- This may be maximized with respect to θ to give the maximum likelihood estimates.

Review: Maximum Likelihood Estimation

- Assume a particular model with unknown parameters, θ
- We can then define the probability of observing a given event conditional on a particular set of parameters.
- We have observed a set of outcomes in the real world.
- It is then possible to choose a set of parameters which are most likely to have produced the observed results.

$$\hat{\theta} = argmax_{\theta} P(X_1, \dots, X_{\mathrm{n}} | \theta)$$

Likelihood

- This is maximum likelihood.

$$\log\big(L(\theta)\big) = \sum_{i=1}^{n} \log(P(X_i | \theta))$$

Log-Likelihood

- It's often both consistent and efficient.
- It provides a standard to compare other estimation techniques.

MLE for Logistic Regression Training

- Training set: $(x_i, y_i), i = 1, \dots, n$

$$l(\beta) = \sum_{i=1}^{n} \{\log Pr(Y = y_i | X = x_i)\}$$
$$= \sum_{i=1}^{n} \{y_i \log Pr(Y = 1 | X = x_i) + (1 - y_i) \log Pr(Y = 0 | X = x_i)\}$$
$$= \sum_{i=1}^{n} \{y_i \log \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} + (1 - y_i) \log \frac{1}{1 + e^{\beta^T x_i}}\}$$
$$= \sum_{i=1}^{n} \{y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})\}$$

Extra

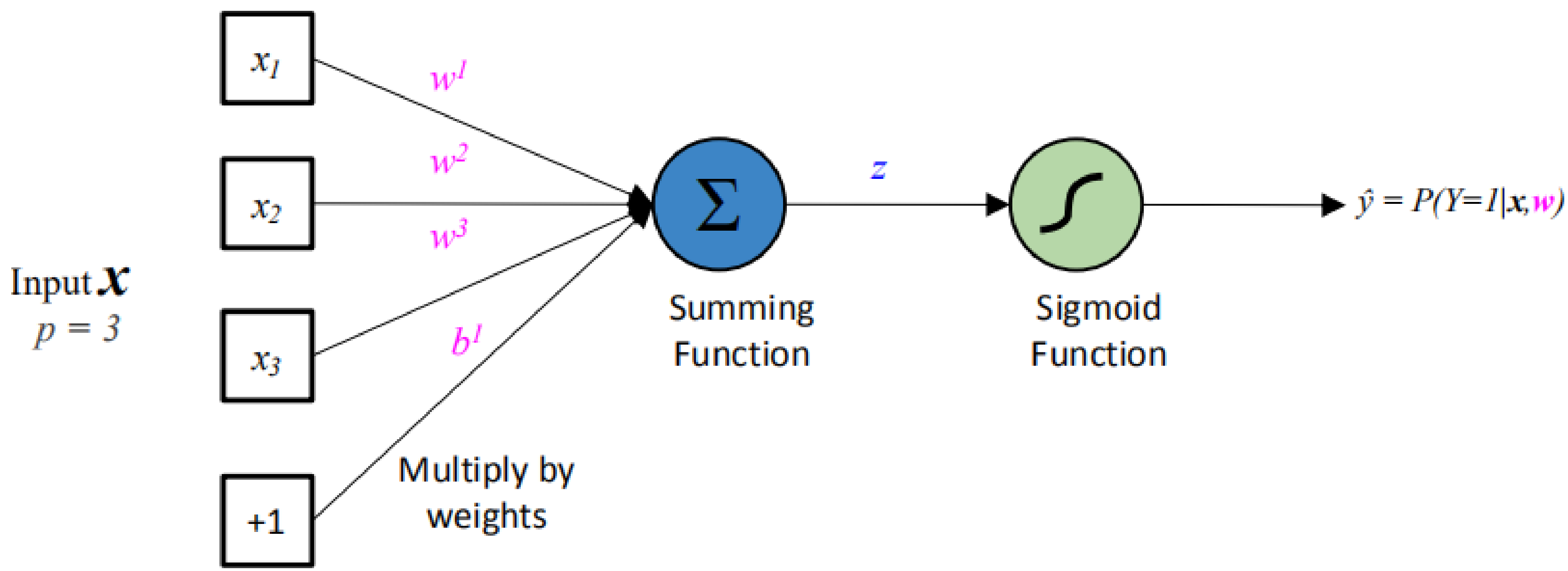
- Rewrite Logistic Regression as two stages:
  - First: summing

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

  - Second: Sigmoid Squashing

$$\hat{y} = P(y = 1|x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}} = \frac{e^z}{1 + e^z}$$

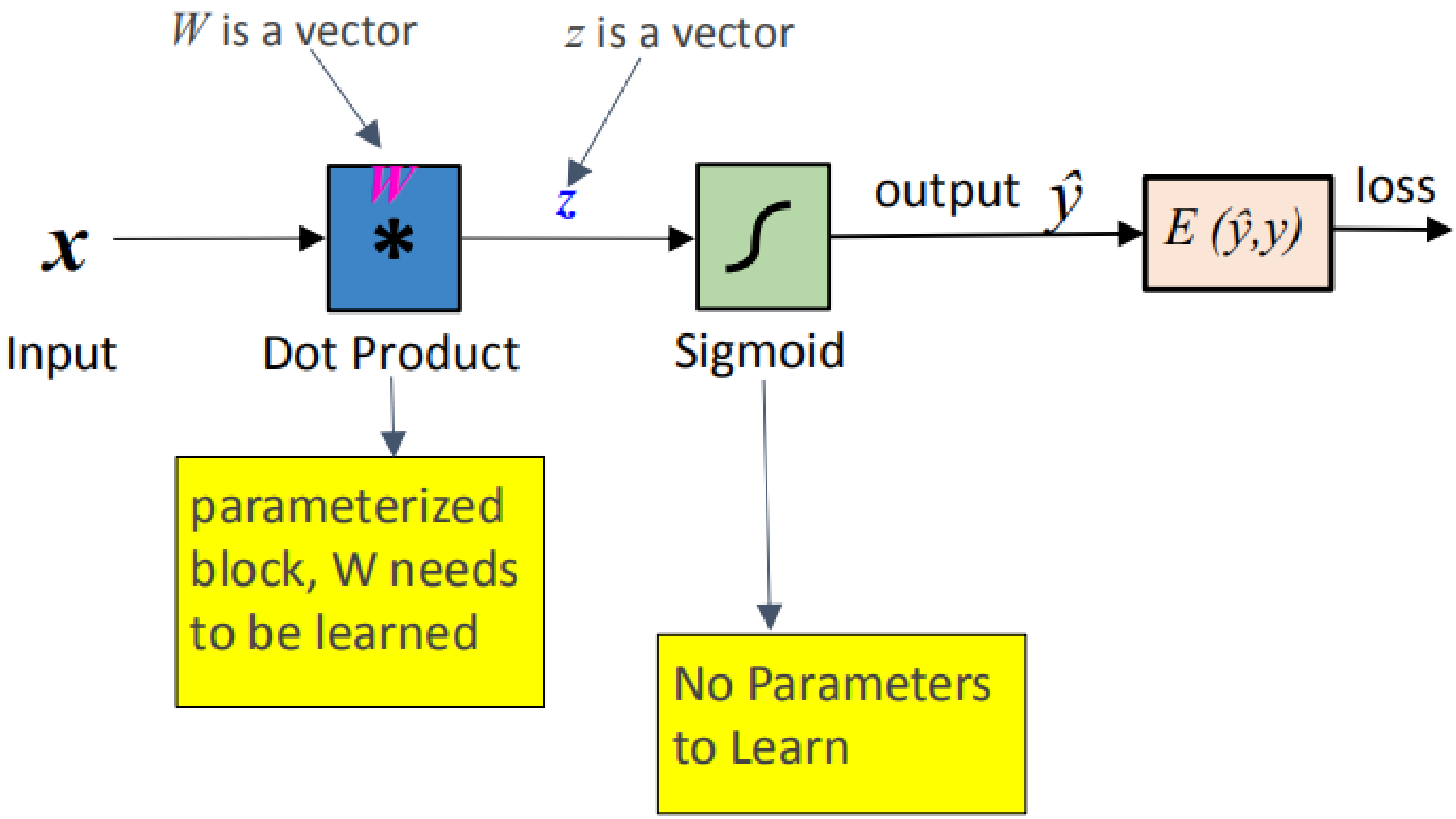
“Neuron”: Block View of Logistic Regression



$$z = w^T \cdot \mathbf{x} + b$$

$$y = sigmoid(z) = \frac{e^z}{1 + e^z}$$

e.g., "Block View" of Logistic Regression



Three major sections for classification

- Discriminative
  - directly estimate a decision rule/boundary
  - e.g., support vector machine, decision tree, logistic regression
  - e.g. neural networks (NN), deep NN
- Generative
  - build a generative statistical model
  - e.g., Bayesian networks, Naïve Bayes classifier
- Instance based classifiers
  - Use observation directly (no models)
  - e.g. K nearest neighbors

References

- https://qiyanjun.github.io/2019f-UVA-CS6316-MachineLearning/
- Prof. Tan, Steinbach, Kumar's "Introducoon to Data Mining" slide
- Prof. Andrew Moore's slides
- Prof. Eric Xing's slides
- Prof. Ke Chen NB slides
- Hasoe, Trevor, et al. *The elements of sta8s8cal learning*. Vol. 2. No. 1. New York: Springer, 2009.

Thanks for listening