UVA CS 6316/4501 – Fall 2016 Machine Learning

Lecture 14: Logistic Regression / Generative vs. Discriminative

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

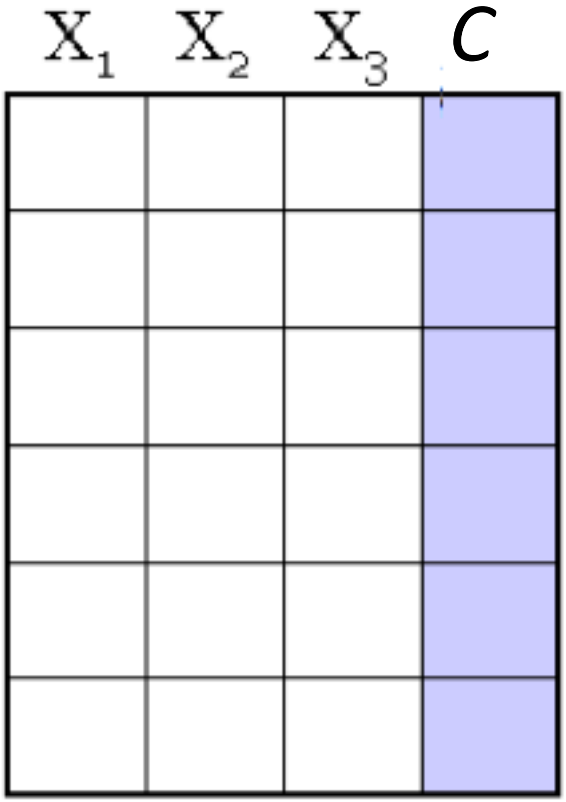
Where are we ? → Three major sections for classification

- 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



A Dataset for classification

$$f : X \longrightarrow C$$

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

Generative: $$\arg\max_C P(C \mid X) = \arg\max_C P(X, C) = \arg\max_C P(X \mid C) P(C)$$

Discriminative: $$\arg\max_C P(C / \mathbf{X}) \quad C = c_1, \cdots, c_L$$

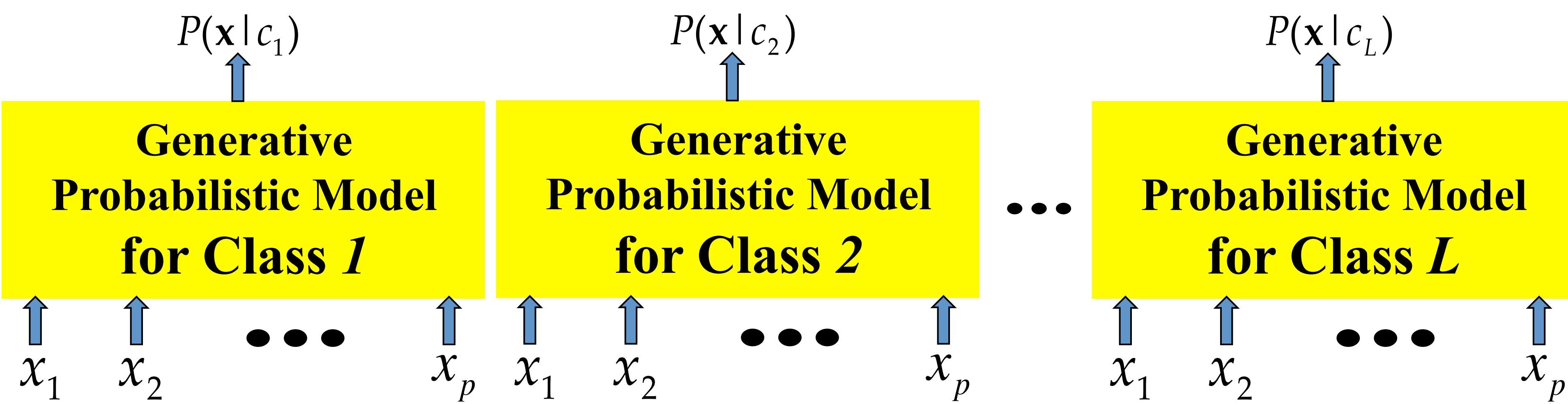
- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

Establishing a probabilistic model for classification (cont.)

- **(1) Generative model**

$$\arg\max_{C} P(C \mid X) = \arg\max_{C} P(X, C)$$

$$= \arg\max_{C} P(X \mid C) P(C)$$



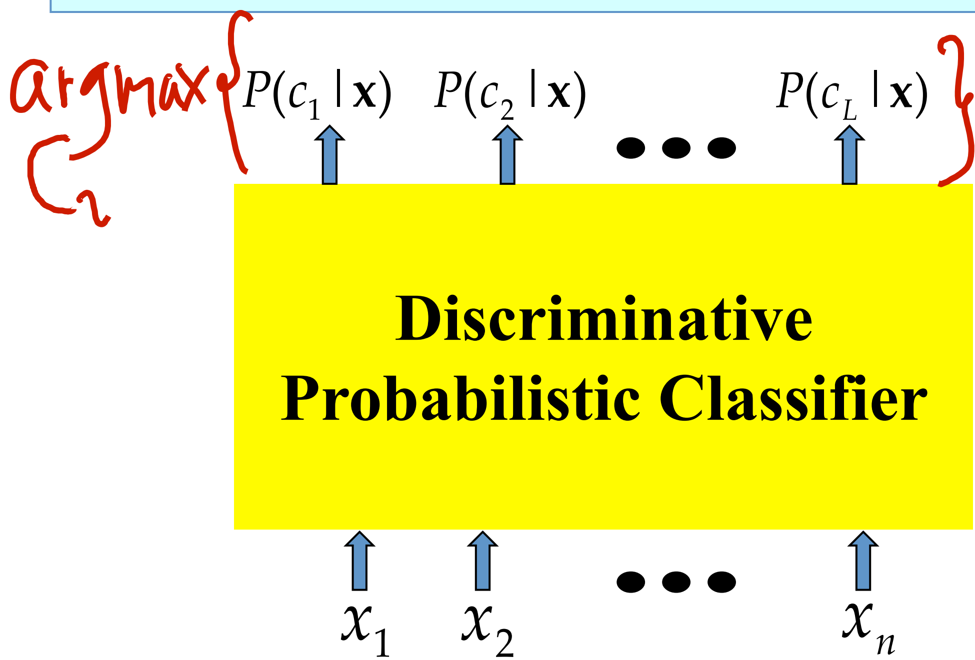
$$\mathbf{x} = (x_1, x_2, \cdots, x_p)$$

Adapt from Prof. Ke Chen NB Slides

Establishing a probabilistic model for classification

– **(2) Discriminative model**

$$P(C\,|\,\mathbf{X}) \quad C = c_1, \cdots, c_L,\ \mathbf{X} = (X_1, \cdots, X_n)$$



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

Adapt from Prof. Ke Chen NB slides

Today : Generative vs. Discriminative

- ✓ Why Bayes Classification – MAP Rule?
  - Empirical Prediction Error
  - 0-1 Loss function for Bayes Classifier
- ✓ Logistic regression
- ✓ Generative vs. Discriminative

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

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

WHY ?

MAP = Maximum Aposteriori Probability

Adapt From Carols' prob tutorial

0-1 LOSS for Classification

- Procedure for categorical output variable C

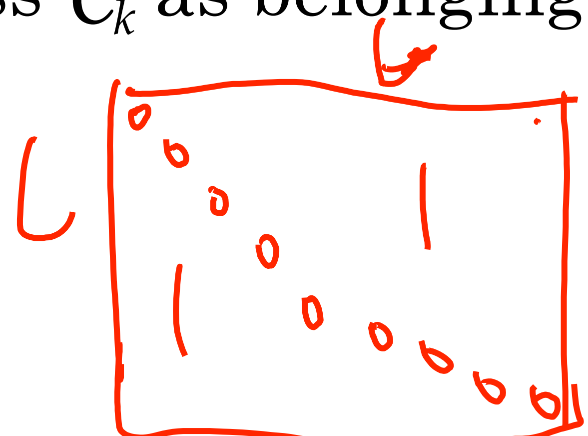
if $k = \ell$, $L(k, \ell) = 0$

- Frequently, 0-1 loss function used: $L(k, \ell)$

if $k \neq \ell$, $L(k, \ell) = 1$

- $L(k, \ell)$ is the price paid for misclassifying an element from class C_k as belonging to class C_ℓ

➔ *L*L matrix*



$C_1, C_2, \cdots, C_L$

Expected prediction error (EPE)

- Expected prediction error (EPE), with expectation taken w.r.t. the joint distribution Pr(C,X)
  - $\Pr(C,X)=\Pr(C\mid X)\Pr(X)$

e.g. 0-1 loss

$E_x(X)$

$E_x(g(x))$

$$\mathrm{EPE}(f)=E_{X,C}(L(C,f(X))$$

$$=\mathrm{E}_X\sum_{k=1}^{L}L[C_k,f(X)]\Pr(C_k\mid X)$$

Consider sample population distribution

$$EPE(f) = E_{X,C}\left(L(C, f(X))\right)$$

$$= E_X E_{C|X}\left[L(C, f(X)) \,|\, X\right]$$

$$E_C(C) = \sum_{i=1}^{L} C_i P(C_i)$$

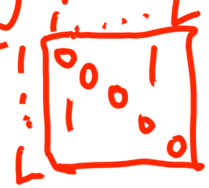
$C|X$: Discrete RV's Expectation

$$= E_X \sum_{k=1}^{L} L[C_k, f(X)] \Pr(C_k | X)$$

$$\underset{f}{\operatorname{argmin}}\ EPE(f(X))$$

$\Rightarrow$ Pointwise minization when $X = x$

$$\Rightarrow \hat{f}(X = x) = \underset{f(x) \in C}{\operatorname{argmin}} \sum_{k=1}^{L} L[C_k, f(x)] \Pr(C_k | X = x)$$



$$\Rightarrow \hat{f}(x) = \underset{C_k \in \{C_1, C_2, C_3, \ldots, C_L\}}{\operatorname{argmax}} \Pr(C_k | X = x)$$

$$\begin{cases} p(C_1 | x) \\ p(C_2 | x) \\ \vdots \\ p(C_L | x) \end{cases}$$

Expected prediction error (EPE)

$$\mathrm{EPE}(f) = E_{X,C}(L(C, f(X)) = \mathrm{E}_X \sum_{k=1}^{K} L[C_k, f(X)]\Pr(C_k \mid X)$$

Consider sample population distribution

- Pointwise minimization suffices

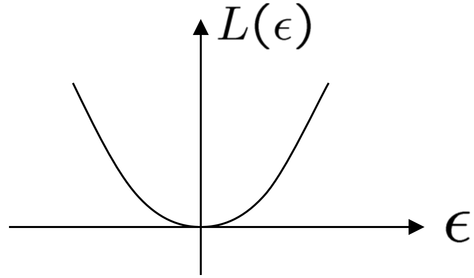
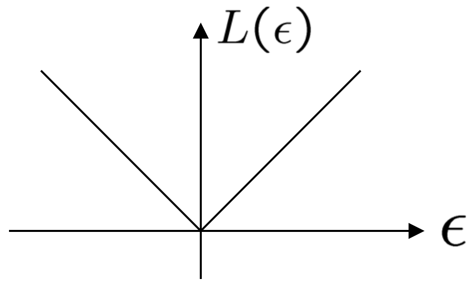
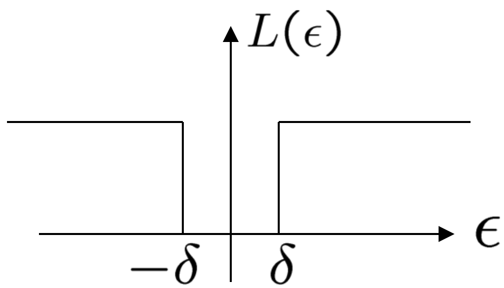
- ➔ simply $$\hat{f}(X) = \operatorname{argmin}_{g \in C} \sum_{k=1}^{K} L(C_k, g)\Pr(C_k \mid X = x)$$

Bayes Classifier

$$\hat{f}(X) = C_k \text{ if}$$

$$\Pr(C_k \mid X = x) = \max_{g \in C} \Pr(g \mid X = x)$$

SUMMARY: WHEN EPE USES DIFFERENT LOSS

Loss Function		Estimator $\hat{f}(x)$
L_2		$EPE = E_{X,Y}(Y - f(X))^2$ $\hat{f}(x) = E[Y\|X = x]$
L_1		$\hat{f}(x) = \text{median}(Y\|X = x)$
0-1		$\hat{f}(x) = \arg\max_Y P(Y\|X = x)$ (Bayes classifier / MAP)

Today : Generative vs. Discriminative

- ✓ Why Bayes Classification – MAP Rule?
  - Empirical Prediction Error
  - 0-1 Loss function for Bayes Classifier
- ✓ Logistic regression
- ✓ Generative vs. Discriminative

Multivariate linear regression to Logistic Regression

$$\mathbf{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

linear

Dependent	Independent variables
Predicted	Predictor variables
Response variable	Explanatory variables
Outcome variable	Covariables

Logistic regression for binary classification

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

logistic

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

(1) Linear decision boundary [Seperate two classes]

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

linear hyperplane

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

⇑

Boundary points

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

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

(1) Linear decision boundary [Seperate two classes]

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

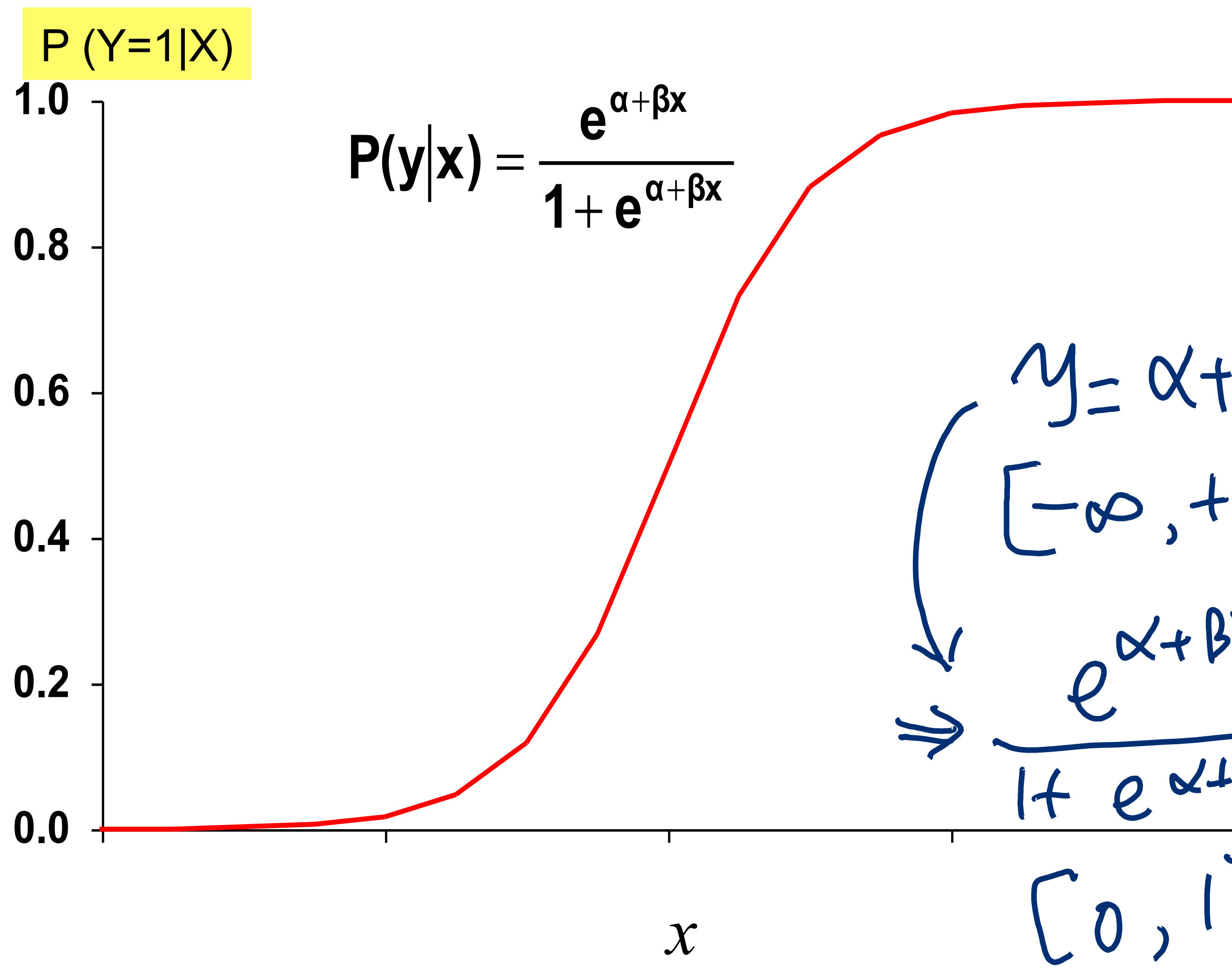
(2) p(y|x) $\Rightarrow$

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

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

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

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.

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.

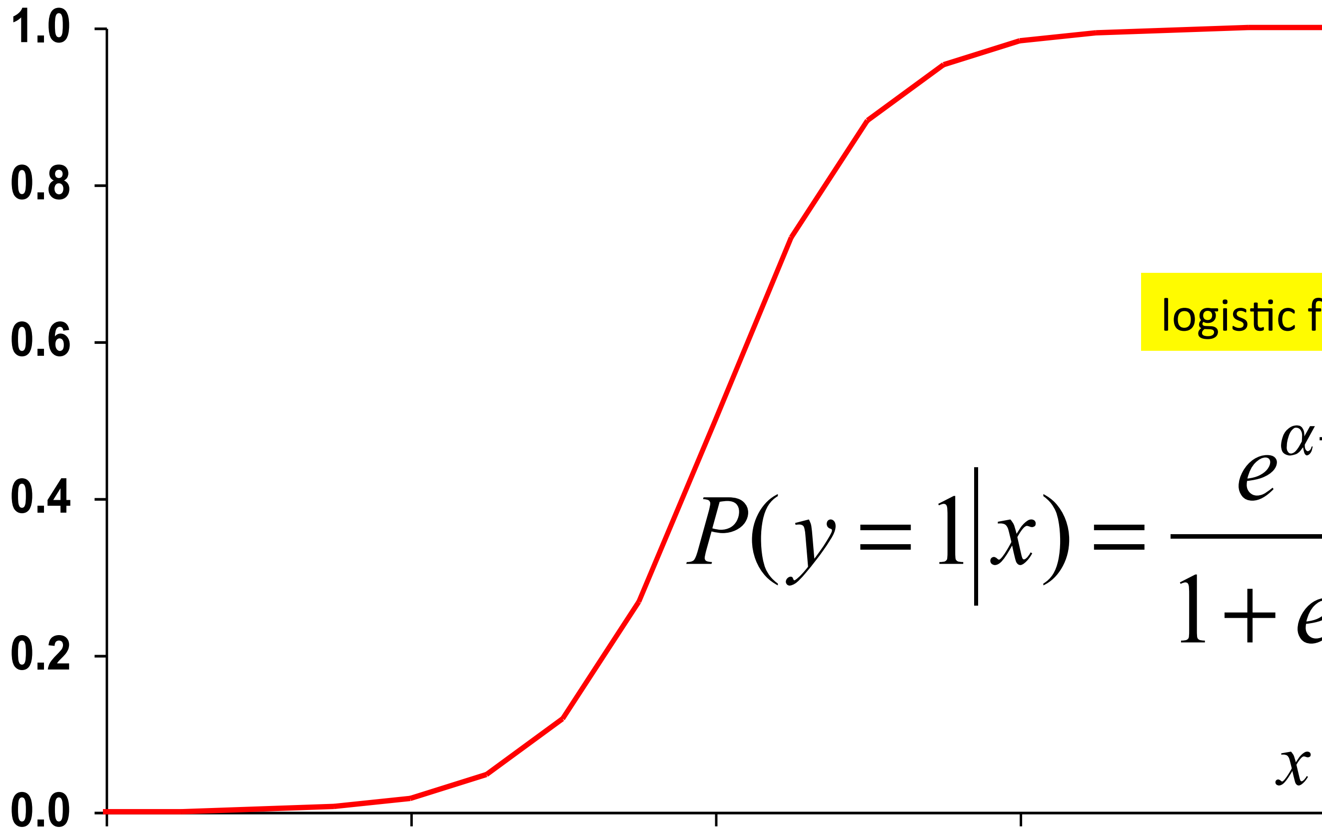
The main interest ➔ predicting the probability that an event occurs (i.e., the probability that p(y=1|x)).

Discriminative

Logistic regression models for binary target variable coded 0/1.

P (y=1|X)

e.g. Probability of disease



logistic function

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

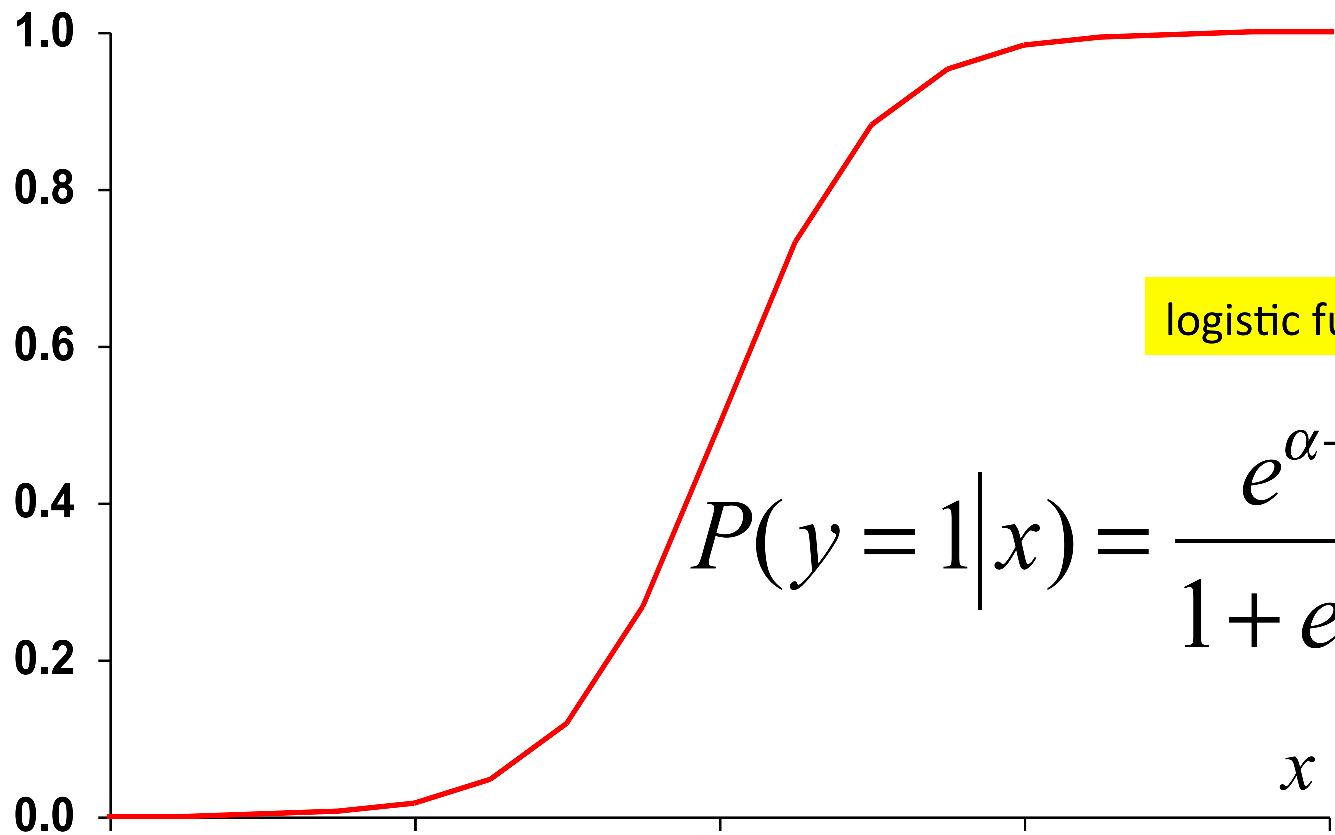
Logit function

Discriminative

Logistic regression models for binary target variable coded 0/1.

P (y=1|X)

e.g. Probability of disease



logistic function

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

Logit function

Decision Boundary → equals to zero

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

The logistic function (2)

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

logistic

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

logit / log-odd

Logit of $P(y|x)$

The logistic function (3)

- Advantages of the logit
  - Simple transformation of P(y|x)
  - Linear relationship with x
  - Can be continuous (Logit between - inf to + infinity)
  - Directly related to the notion of log odds of target event

$$z = \log\left(\frac{p}{1-p}\right)$$

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta x \qquad \frac{P}{1-P} = e^{\alpha+\beta x}$$

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.

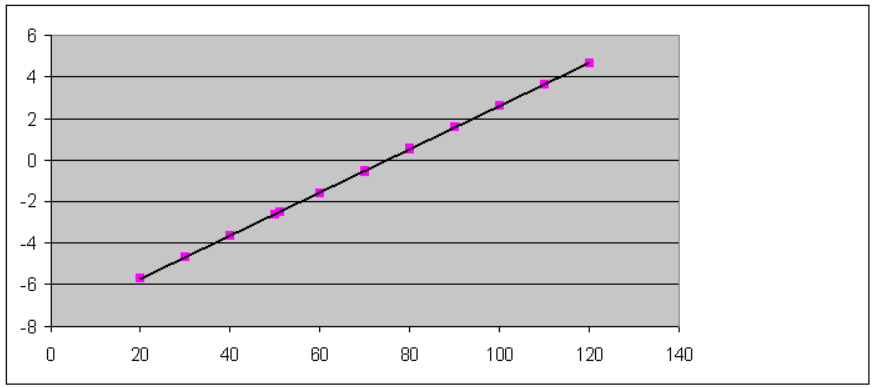
Binary Logistic Regression (K=2)

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

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

ln[p/(1-p)]

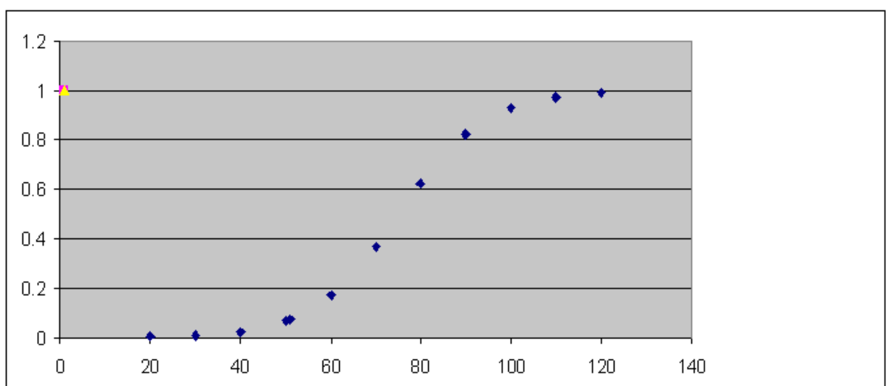
Odds= p/(1-p)



x

- Logistic Distribution

P (Y=1|x)

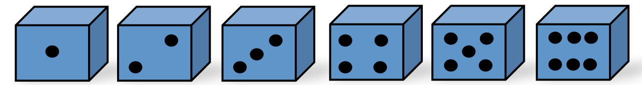


x

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

p 1-p

Binary → Multinomial Logistic Regression Model



p(G | x)

Directly models the posterior probabilities as the output of regression

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

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

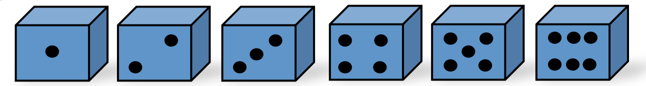
x is p-dimensional input vector

\beta$_k$ is a p-dimensional vector for each k

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

Note that the class boundaries are linear

Binary → Multinomial Logistic Regression Model



Directly models the posterior probabilities as the output of regression

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

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

x is p-dimensional input vector

\beta$_k$ is a p-dimensional vector for each k

Total number of parameters is $(K-1)(p+1)$

Note that the class boundaries are linear

Today : Generative vs. Discriminative

- ✓ Why Bayes Classification – MAP Rule?
  - Empirical Prediction Error
  - 0-1 Loss function for Bayes Classifier
- ✓ Logistic regression
  - Parameter Estimation for LR
- ✓ Generative vs. Discriminative

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

Parameter Estimation for LR → MLE from the data

- RECAP: Linear regression → Least squares

- Logistic regression: → Maximum likelihood estimation

MLE for Logistic Regression Training

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

For Bernoulli distribution

$$p(y \mid x)^y (1-p)^{1-y}$$

Training set: (x_i, y_i), i=1,…,N

(conditional) Log-likelihood:

How?

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

$p(y_i|x_i)$

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

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

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

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*

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

y_i

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

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

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

Newton-Raphson for LR (optional)

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} (y_i - \frac{\exp(\beta^T x)}{1+\exp(\beta^T x)}) x_i = 0$$

Vector

(*p*+1) Non-linear equations to solve for (*p*+1) unknowns β

Solve by Newton-Raphson method:

$$\beta^{new} \leftarrow \beta^{old} - [(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T})]^{-1} \frac{\partial l(\beta)}{\partial \beta},$$

$$\text{where, } (\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}) = -\sum_{i=1}^{N} x_i x_i^T (\frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)})(\frac{1}{1+\exp(\beta^T x_i)})$$

$p(x_i ; \beta)$

$1 - p(x_i ; \beta)$

minimizes a quadratic approximation
to the function we are really interested in.

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mathbf{H}_K^{-1} \mathbf{g}_k$$

Newton-Raphson for LR…

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N}\left(y_i - \frac{\exp(\beta^T x)}{1+\exp(\beta^T x)}\right)x_i = X^T(y-p)$$

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

$$\left(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}\right) = -X^T W X$$

So, NR rule becomes: $$\beta^{new} \leftarrow \beta^{old} + (X^T W X)^{-1} X^T (y-p),$$

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}_{N-by-(p+1)}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N-by-1}, \quad p = \begin{bmatrix} \exp(\beta^T x_1)/(1+\exp(\beta^T x_1)) \\ \exp(\beta^T x_2)/(1+\exp(\beta^T x_2)) \\ \vdots \\ \exp(\beta^T x_N)/(1+\exp(\beta^T x_N)) \end{bmatrix}_{N-by-1},$$

$X : N\times(p+1)$ matrix of x_i

$y : N\times 1$ matrix of y_i

$p : N\times 1$ matrix of $p(x_i;\beta^{old})$

$W : N\times N$ diagonal matrix of $p(x_i;\beta^{old})(1-p(x_i;\beta^{old}))$

$$\left(\frac{\exp(\beta^T x_i)}{(1+\exp(\beta^T x_i))}\right)\left(1-\frac{1}{(1+\exp(\beta^T x_i))}\right)$$

Newton-Raphson for LR…

- Newton-Raphson
  - $\beta^{new} = \beta^{old} + (X^T W X)^{-1} X^T (y - p)$

    $= (X^T W X)^{-1} X^T W (X\beta^{old} + W^{-1}(y - p))$

    $= (X^T W X)^{-1} X^T W z$

Re expressing Newton step as weighted least square step

$(X^T X)^{-1} X^T Y$ (W, Wz)

  - Adjusted response

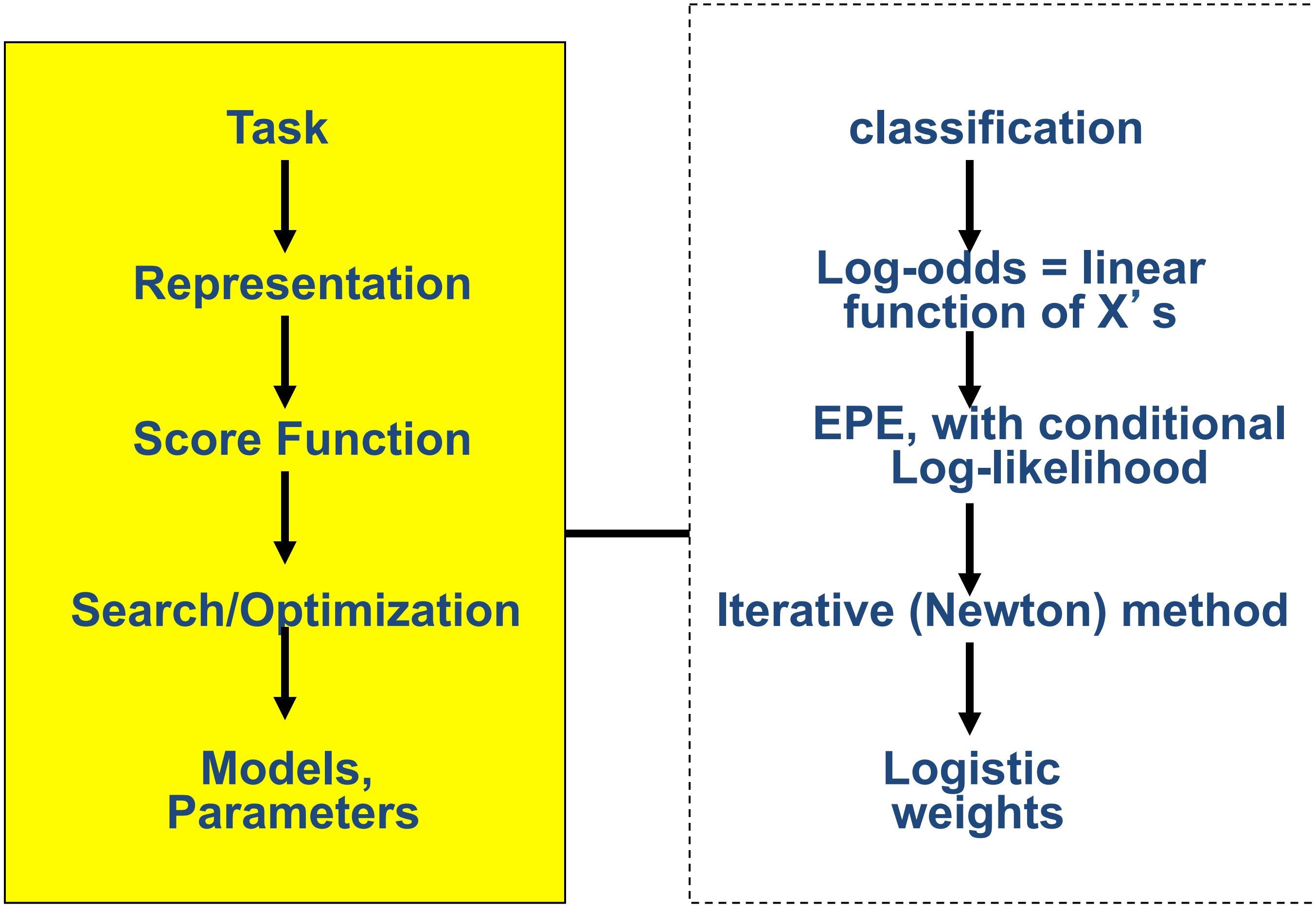
    $z = X\beta^{old} + W^{-1}(y - p)$

  - Iteratively reweighted least squares (IRLS)

$$\beta^{new} \leftarrow \arg\min_{\beta} (z - X\beta^T)^T W (z - X\beta^T)$$

$$\leftarrow \arg\min_{\beta} (y - p)^T W^{-1} (y - p)$$

Logistic Regression



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

Today : Generative vs. Discriminative

- ✓ Why Bayes Classification – MAP Rule?
  - Empirical Prediction Error
  - 0-1 Loss function for Bayes Classifier
- ✓ Logistic regression
- ✓ Generative vs. Discriminative

Discriminative vs. Generative

Generative approach

- Model the joint distribution p(X, C) using

  $p(X \mid C = c_k)$ and $p(C = c_k)$

Class prior

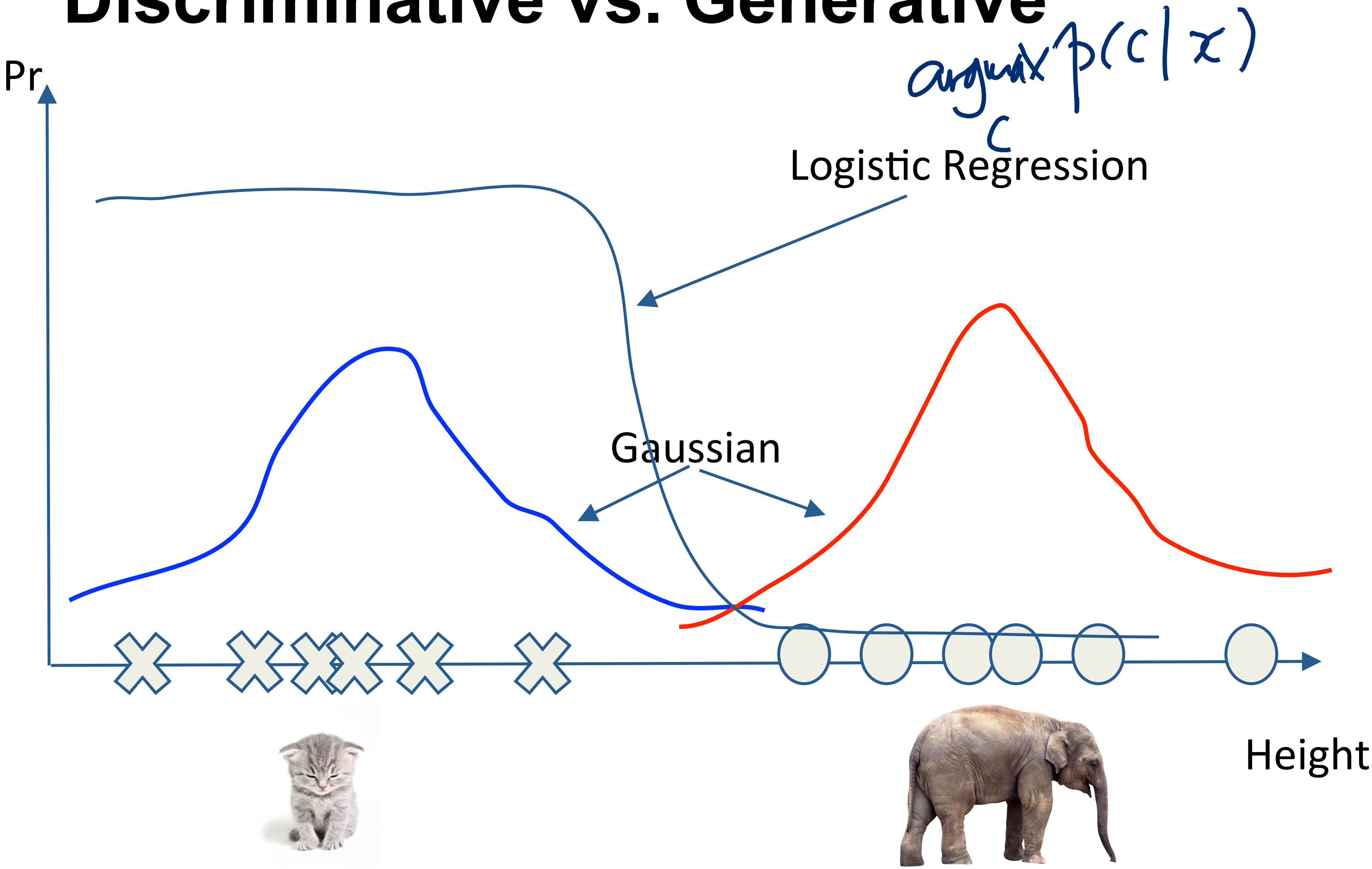
Discriminative approach

- Model the conditional distribution p(c| X) directly

e.g.,

$$p(C=1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * X)}}$$

Discriminative vs. Generative



LDA vs. Logistic Regression

$\Rightarrow$ mean $Kp + p^2$ conv

- **LDA (Generative model)**
  - Assumes Gaussian class-conditional densities and a common covariance
  - Model parameters are estimated by maximizing the full log likelihood, parameters for each class are estimated independently of other classes, $Kp + \frac{p(p+1)}{2} + (K-1)$ parameters
  - Makes use of marginal density information $\Pr(x)$
  - Easier to train, low variance, more efficient if model is correct
  - Higher asymptotic error, but converges faster

$\Rightarrow (K-1)(p+1)$

- **Logistic Regression (Discriminative model)**
  - Assumes class-conditional densities are members of the (same) exponential family distribution
  - Model parameters are estimated by maximizing the conditional log likelihood, simultaneous consideration of all other classes, $(K-1)(p+1)$parameters
  - Ignores marginal density information $\Pr(x)$
  - Harder to train, robust to uncertainty about the data generation process
  - Lower asymptotic error, but converges more slowly

Discriminative vs. Generative

- Definitions
  - h_{gen} and h_{dis}: generative and discriminative classifiers
  - $h_{gen, inf}$ and $h_{dis, inf}$: same classifiers but trained on the entire population (asymptotic classifiers)
  - $n \rightarrow$ infinity, $h_{gen} \rightarrow h_{gen, inf}$ and $h_{dis} \rightarrow h_{dis, inf}$

Ng, Jordan,. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." *Advances in neural information processing systems* 14 (2002): 841.

Discriminative vs. Generative

Proposition 1: h_{true}

$$\epsilon\left(h_{dis,\inf}\right) \leq \epsilon\left(h_{gen,\inf}\right)$$

Proposition 1 states that aymptotically, the error of the discriminative logistic regression is smaller than that of the generative naive Bayes. This is easily shown

- p : number of dimensions
- n : number of observations
- ϵ : generalization error

Logistic Regression vs. NBC

Discriminative classifier (Logistic Regression)

- Smaller asymptotic error
- Slow convergence ~ O(p)

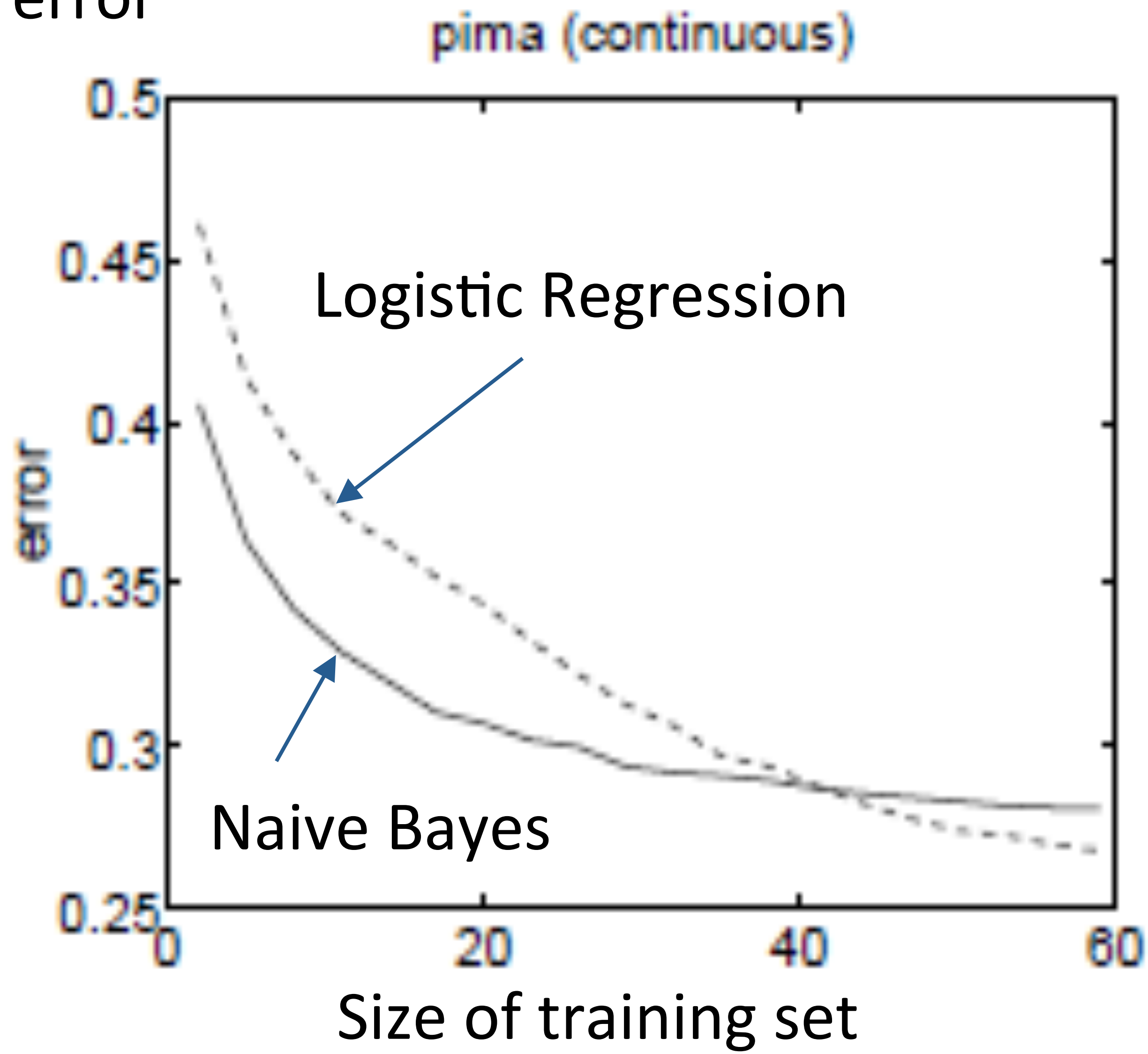
Generative classifier (Naive Bayes)

- Larger asymptotic error
- Can handle missing data (EM)
- Fast convergence ~ O(lg(p))

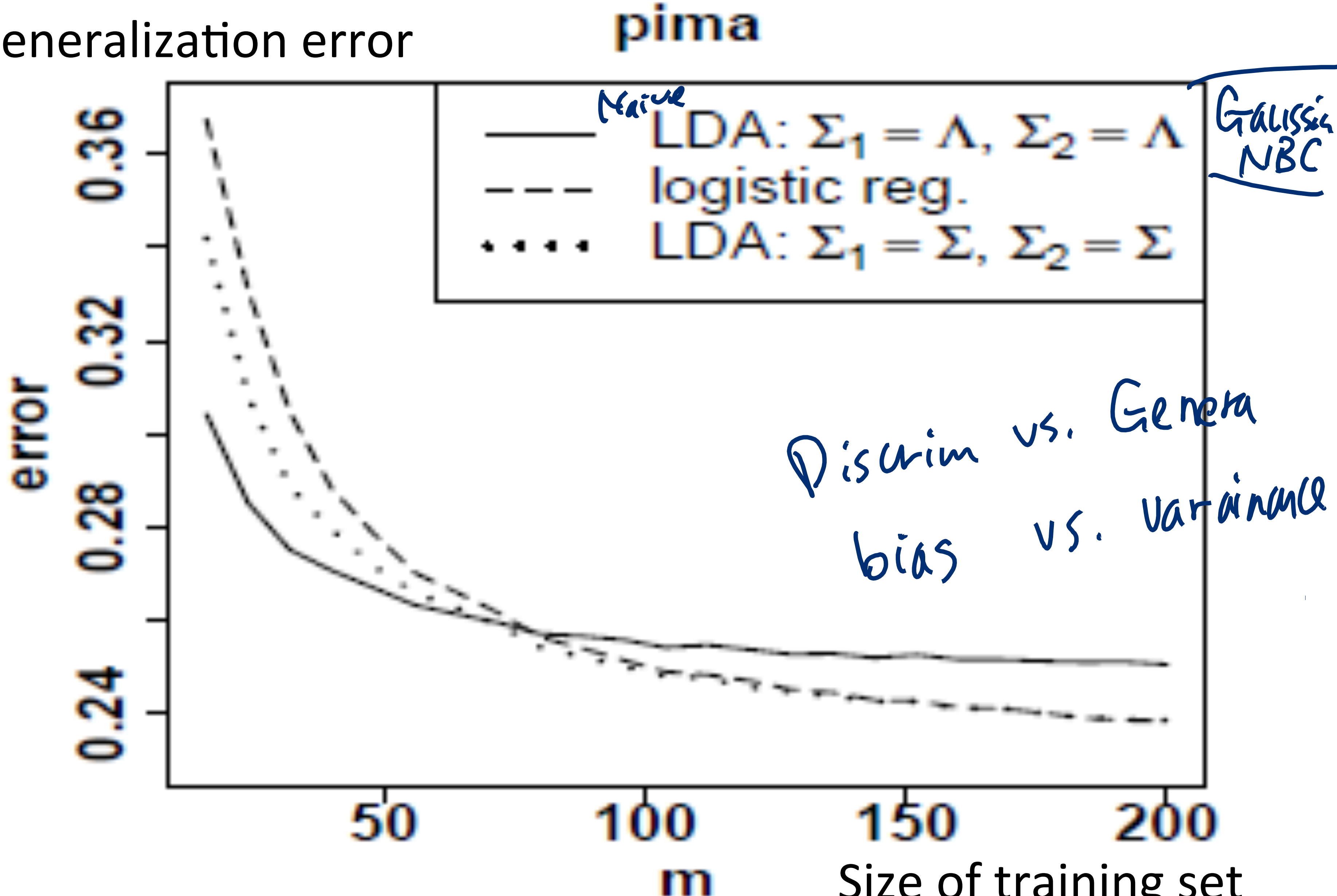
In numerical analysis, the speed at which a convergent sequence approaches its limit is called the rate of convergence.

Ng, Jordan,. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." *Advances in neural information processing systems* 14 (2002): 841.

generalization error



generalization error



Size of training set

Xue, Jing-Hao, and D. Michael Titterington. "Comment on “On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes”."*Neural processing letters* 28.3 (2008): 169-187.

Discriminative vs. Generative

- Empirically, generative classifiers approach their asymptotic error faster than discriminative ones
    - Good for small training set
    - Handle missing data well (EM)
- Empirically, discriminative classifiers have lower asymptotic error than generative ones
    - Good for larger training set

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.