UVA CS 6316 – Fall 2015 Graduate: Machine Learning

Lecture 15: 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

X_1	X_2	X_3	C

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

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

Generative $\quad \underset{C}{\operatorname{argmax}} P(C|X) = \underset{C}{\operatorname{argmax}} P(X,C) = \underset{C}{\operatorname{argmax}} P(X|C)P(C)$

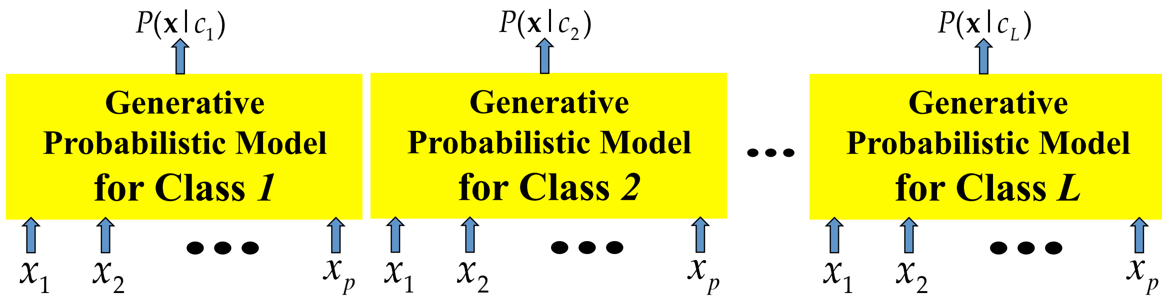
Discriminative $\quad P(C/\mathbf{X}) \quad C = c_1, \cdots, c_L$

- **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**

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



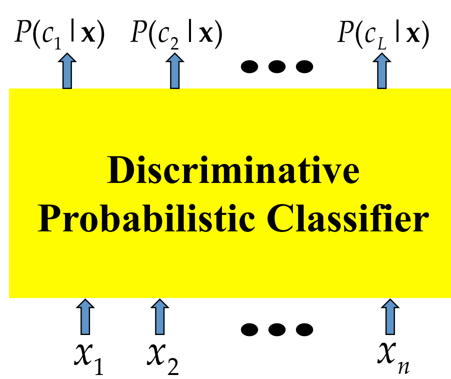
$$\mathbf{x} = (x_1, x_2, \cdots, x_p)$$

Dr. Yanjun Qi / UVA CS 6316 / f15

Establishing a probabilistic model for classification

– **(2) Discriminative model**

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



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

Today :

- ✓ Logistic regression
- ✓ Generative vs. Discriminative

Multivariate linear regression to Logistic Regression

$$\mathbf{y} = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 \mathbf{x}_1 + \boldsymbol{\beta}_2 \mathbf{x}_2 + \ldots + \boldsymbol{\beta}_i \mathbf{x}_i$$

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

Logistic regression for binary classification

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

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

(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) $\Rightarrow$

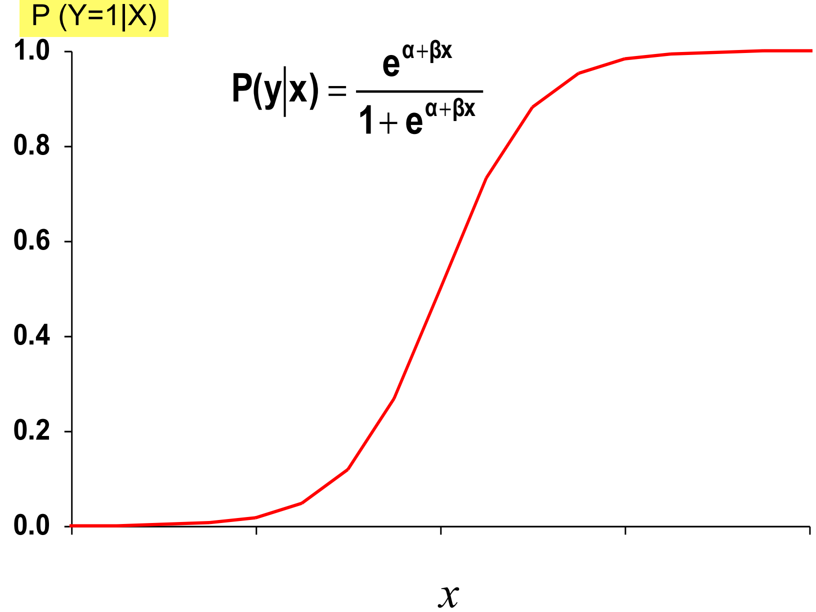
$$\frac{p(y=1|x)}{(1-p(y=1|x))} = e^{\beta^T x}$$

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

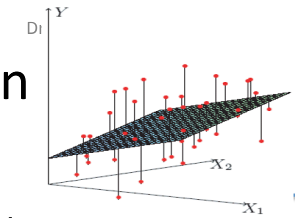
The logistic function (1)
-- is a common "S" shape func

e.g.
Probability of disease

P (Y=1|X)



RECAP: Probabilistic Interpretation of Linear Regression



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

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

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

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

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

- By IID assumption ➔ likelihood ➔ MLE estimator

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

$$l(\theta) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta^T \mathbf{x}_i)^2$$

Logistic Regression—when?

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

Binary

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(y=1|x)
logistic

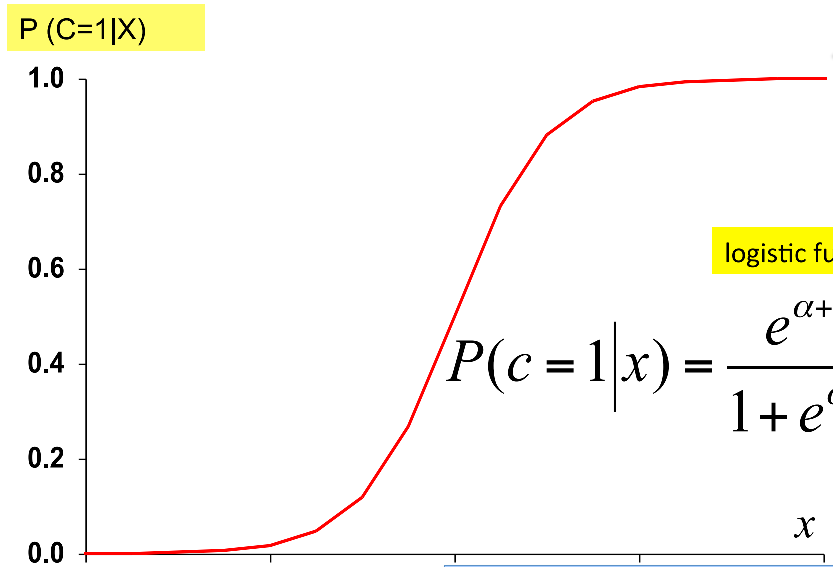
This means we use Bernoulli distribution to model the target variable with its Bernoulli parameter p=p(y=1|x) predefined.

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.

e.g. Probability of disease



logistic function

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

Logit function

Decision Boundary ➔ equals to zero

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

The logistic function (2)

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

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

From probability to logit, i.e. log odds (and back again)

$$z = \log\left(\frac{p}{1-p}\right) \qquad \text{logit / log odd function}$$

$$\frac{p}{1-p} = e^z$$

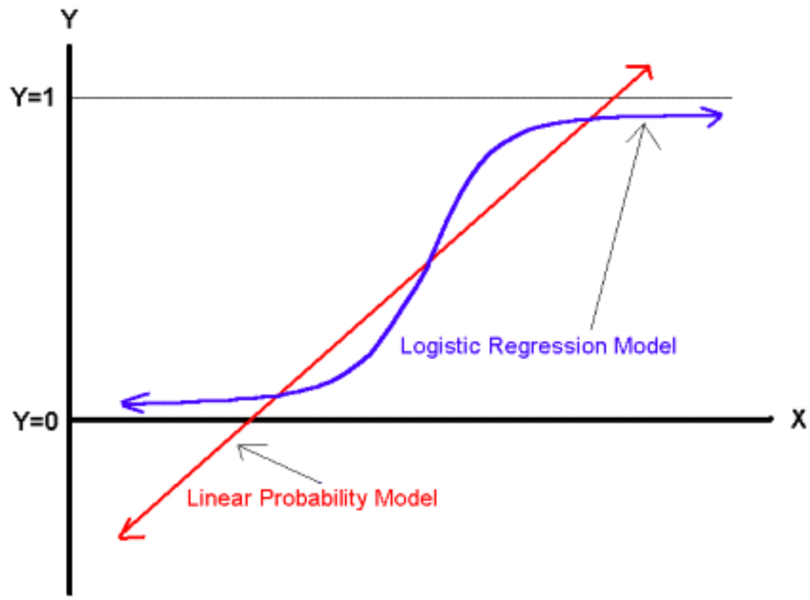
$$p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \qquad \text{logistic function}$$

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

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

Logistic regression – Binary outcome target variable Y



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 / predictors being linearly related to each other.

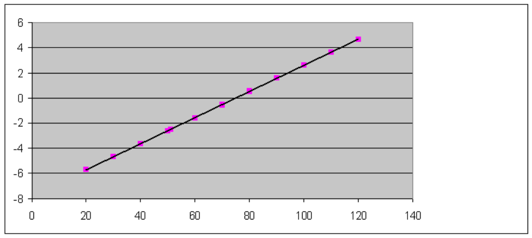
Binary Logistic Regression (K=2)

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

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

ln[p/(1-p)]

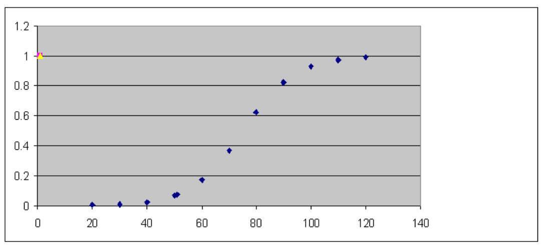
Odds= p/(1-p)



x

- Logistic Distribution

P (Y=1|x)



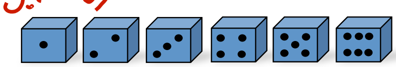
x

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

(e.g. K=6)



Directly models the posterior probabilities as the output of regression

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

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

$p(y=1|x)$

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

$p(y=0|x)$

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

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

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

$\beta_{k0} + \beta_k^T x$

x is p-dimensional input vector

(+1) β_{k0}

\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 :

- ✓ Logistic regression
  - ✓ Parameter estimation
- ✓ Generative vs. Discriminative

Parameter Estimation for LR ➔ MLE from the data

- RECAP: Linear regression ➔ Least squares
- Logistic regression: ➔ Maximum likelihood estimation

RECAP: Probabilistic Interpretation of Linear Regression (cont.)

- Hence the log-likelihood is:

$$l(\theta) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta^T \mathbf{x}_i)^2$$

MLE

- Do you recognize the last term?

Yes it is:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i^T\theta - y_i)^2$$

- Thus under independence assumption, residual square error (RRS) is equivalent to MLE of \theta !

MLE for Logistic Regression Training

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

$$p(y|x)^y(1-p)^{1-y}$$

(conditional) Log-likelihood:

How?

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

$p(y_i|x_i)$

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

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

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

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

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

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

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

Newton-Raphson for LR (optional)

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

Vector

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

Solve by Newton-Raphson method:

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

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

$p(x_i ; \beta)$

$1 - p(x_i ; \beta)$

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…

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

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

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

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 $p(x_i;\beta^{old})(1-p(x_i;\beta^{old}))$

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

Newton-Raphson for LR…

1

- 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

    $(X^T \underset{W}{X})^{-1} X^T \underbrace{Y}_{WZ}$
  - 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

Task	classification
↓	↓
Representation	Log-odds = linear function of X's
↓	↓
Score Function	EPE, with conditional Log-likelihood
↓	↓
Search/Optimization	Iterative (Newton) method
↓	↓
Models, Parameters	Logistic weights

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

Today :

- ✓ 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

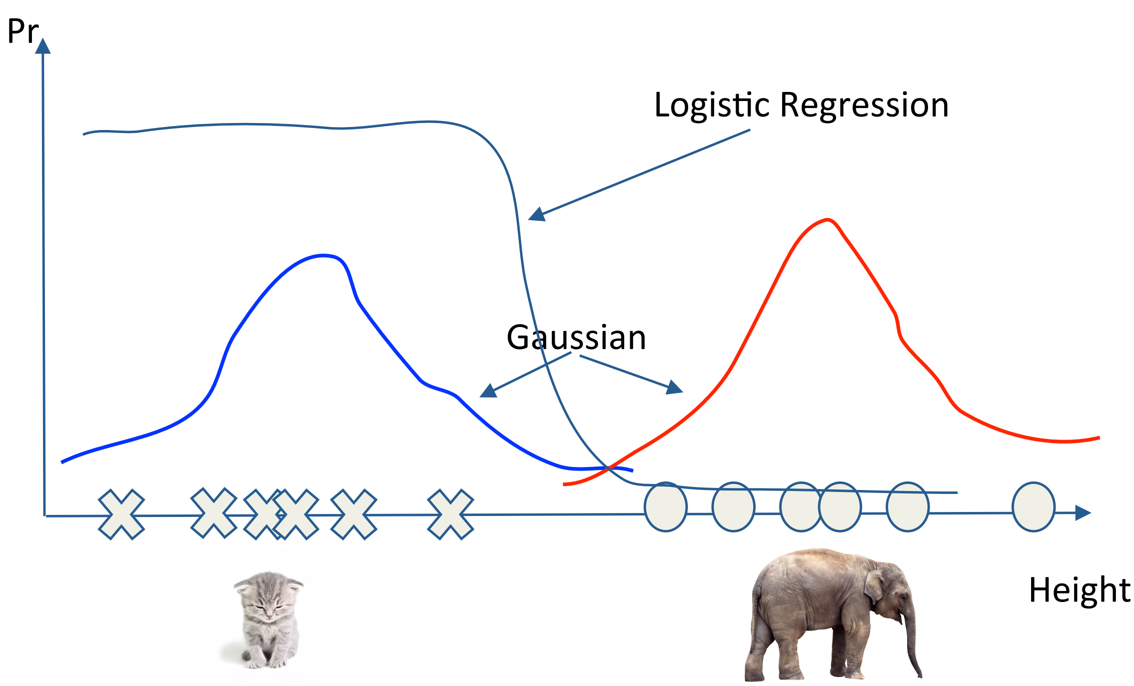
Discriminative approach

- Model the conditional distribution $p(c \mid X)$ directly

e.g.,

$$\frac{1}{1 + e^{-(\beta_0 + \beta_1 * X))}}$$

Discriminative vs. Generative



LDA vs. Logistic Regression

- **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, $K_p + \frac{p(p+1)}{2} + (K-1)$ 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

- **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 → infinity, h_{gen} → $h_{gen, inf}$ and h_{dis} → $h_{dis, inf}$

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:

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

Proposition 1 states that aymptotically, the error of the discriminative logistic re-
gression 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)

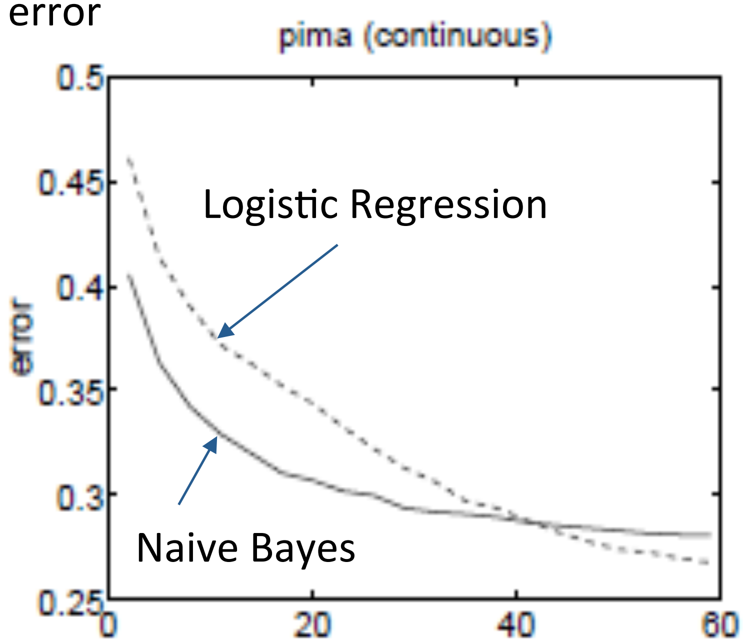
- Smaller asymptotic error
- Slow convergence ~ O(p)

Generative classifier (Naive Bayes)

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

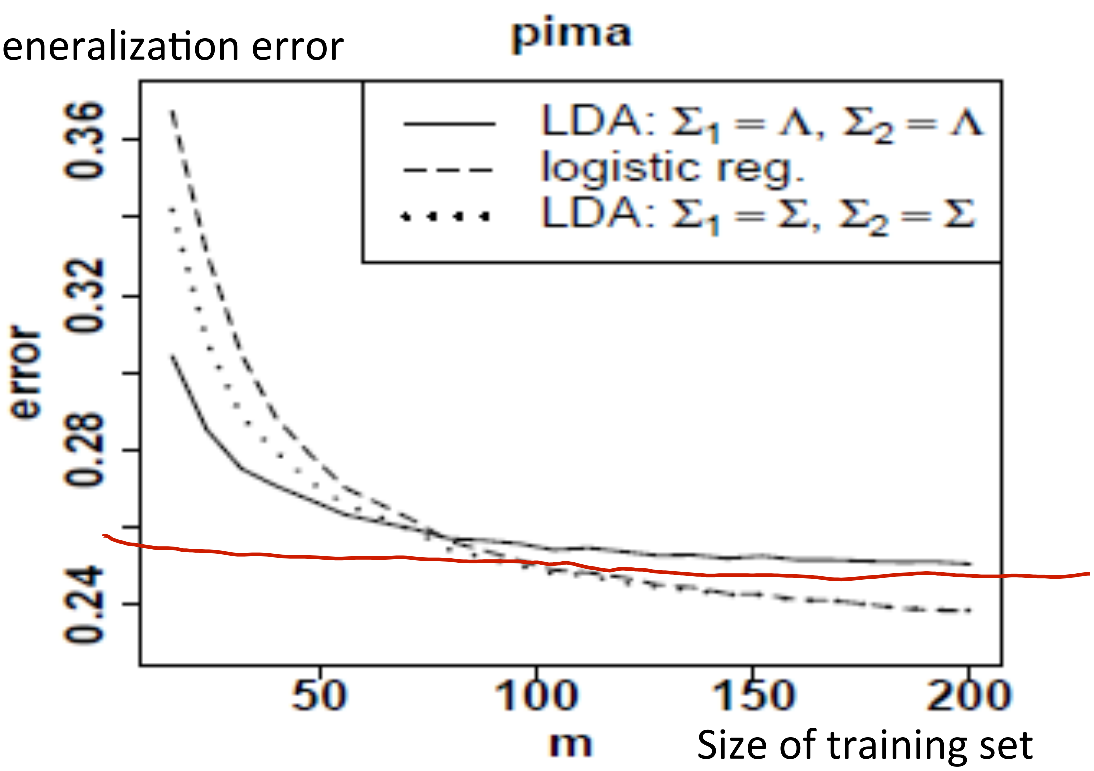
Ng, Jordan,. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." *Advances in neural information processing systems* 14 (2002): 841.

generalization error



Size of training set

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
- Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.