CS145: INTRODUCTION TO DATA MINING

3: Vector Data: Logistic Regression

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

October 8, 2018

Methods to Learn

	Vector Data	Set Data	Sequence Data	Text Data
Classification	**Logistic Regression**; Decision Tree; KNN SVM; NN			Naïve Bayes for Text
Clustering	K-means; hierarchical clustering; DBSCAN; Mixture Models			PLSA
Prediction	**Linear Regression** GLM*			
Frequent Pattern Mining		Apriori; FP growth	GSP; PrefixSpan	
Similarity Search			DTW	

Vector Data: Logistic Regression

- Classification: Basic Concepts
- Logistic Regression Model
- Generalized Linear Model*
- Summary

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Prediction Problems: Classification vs. Numeric Prediction

- Classification
  - predicts categorical class labels
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Numeric Prediction
  - models continuous-valued functions, i.e., predicts unknown or missing values
- Typical applications
  - Medical diagnosis: if a tumor is cancerous or benign
  - Fraud detection: if a transaction is fraudulent
  - Web page categorization: which category it is

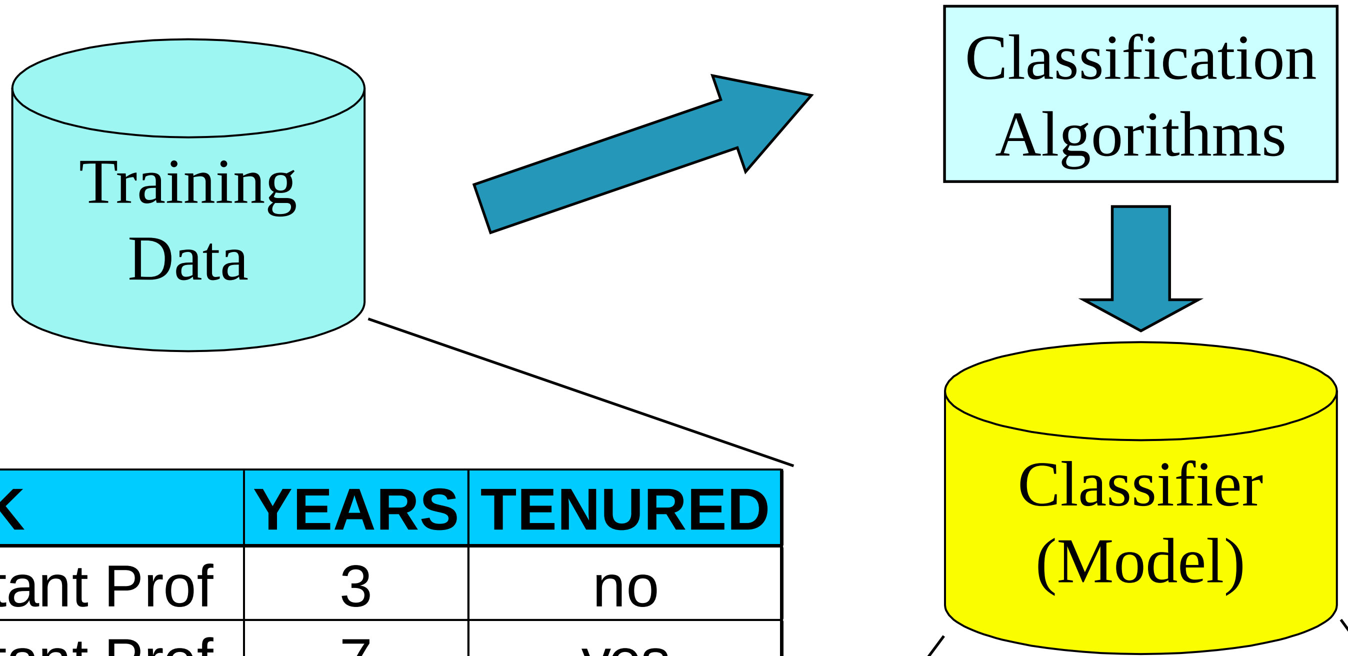
Classification—A Two-Step Process (1)

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
    - For data point *i*: $< \boldsymbol{x_i}, y_i >$
    - Features: $\boldsymbol{x_i}$; class label: y_i
  - The model is represented as classification rules, decision trees, or mathematical formulae
    - Also called classifier
  - The set of tuples used for model construction is training set

Classification—A Two-Step Process (2)

- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Test set is independent of training set (otherwise overfitting)
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
      - Most used for binary classes
  - If the accuracy is acceptable, use the model to classify new data
- Note: If *the test set* is used to select models, it is called validation (test) set

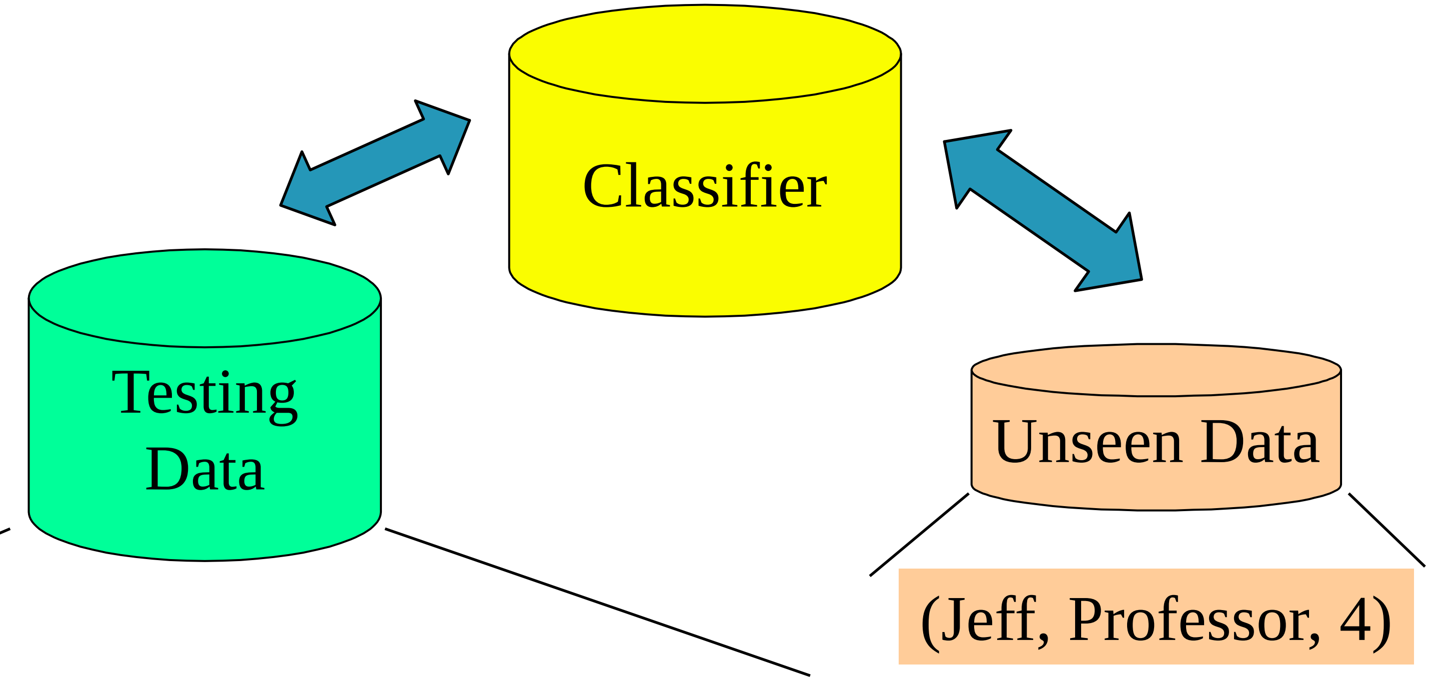
Process (1): Model Construction



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = 'professor'
OR years > 6
THEN tenured = 'yes'

Process (2): Using the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Tenured?

Yes

Vector Data: Logistic Regression

- Classification: Basic Concepts
- Logistic Regression Model
- Generalized Linear Model*
- Summary

Linear Regression VS. Logistic Regression

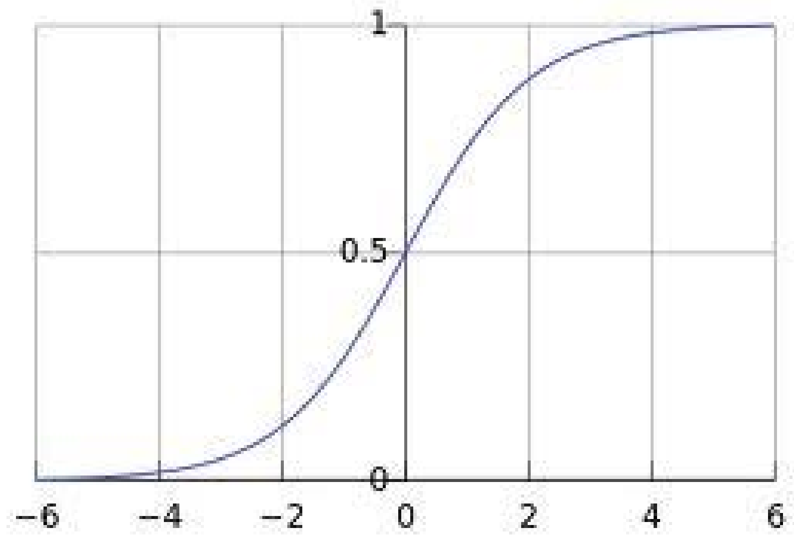
- Linear Regression (prediction)
  - Y: $continuous\ value\ (-\infty, +\infty)$
    - $\mathrm{Y} = \boldsymbol{x}^T \boldsymbol{\beta} = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$
    - $\mathrm{Y}|\boldsymbol{x}, \beta \sim N(\boldsymbol{x}^T\beta, \sigma^2)$

- Logistic Regression (classification)
  - Y: $discrete\ value\ from\ m\ classes$
  - $p(Y = C_j) \in [0,1]\ and\ \sum_j p(Y = C_j) = 1$

Logistic Function

- Logistic Function / sigmoid function:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



$$Note: \sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Modeling Probabilities of Two Classes

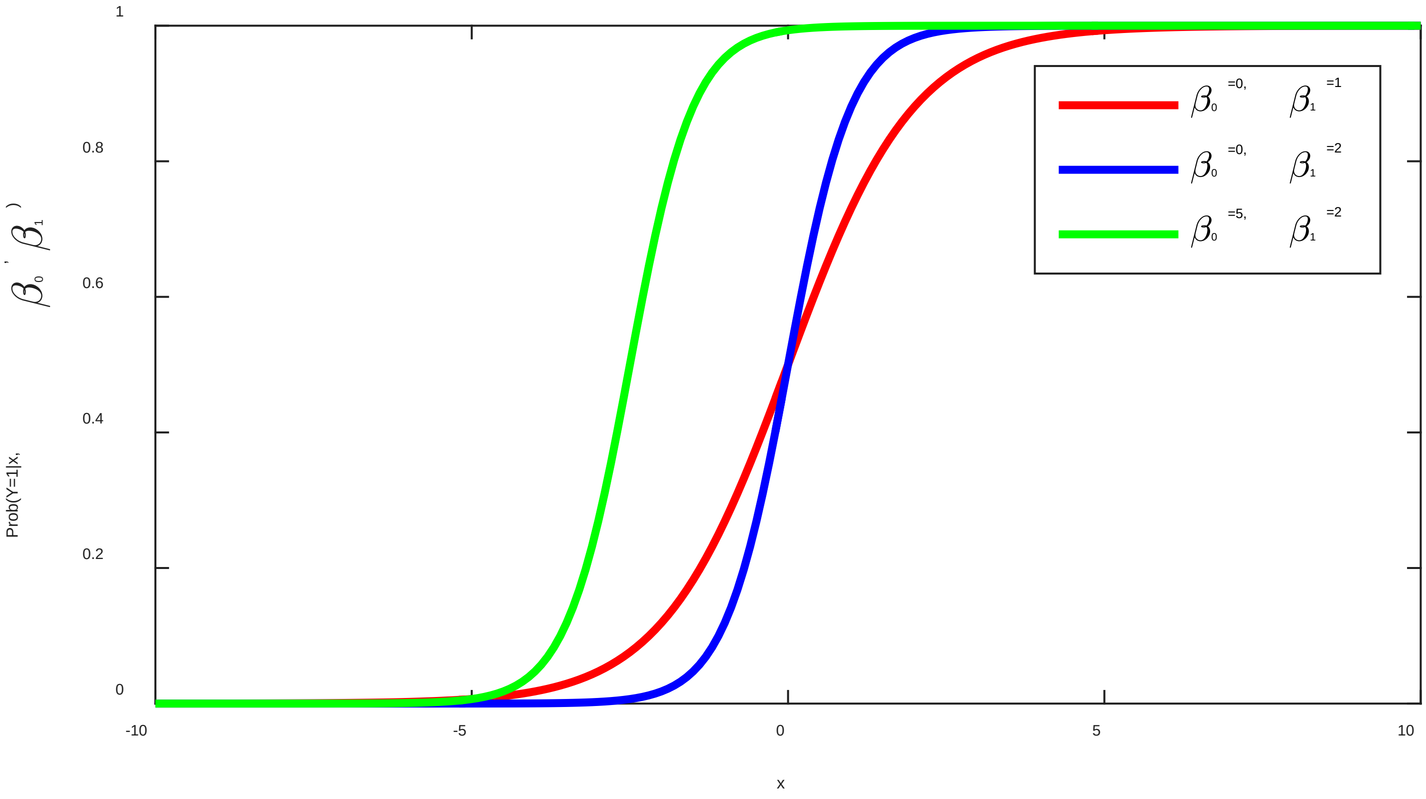
- $P(Y = 1|X, \beta) = \sigma\left(X^T\beta\right) = \frac{1}{1+\exp\{-X^T\beta\}} = \frac{\exp\{X^T\beta\}}{1+\exp\{X^T\beta\}}$
- $P(Y = 0|X, \beta) = 1 - \sigma\left(X^T\beta\right) = \frac{\exp\{-X^T\beta\}}{1+\exp\{-X^T\beta\}} = \frac{1}{1+\exp\{X^T\beta\}}$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

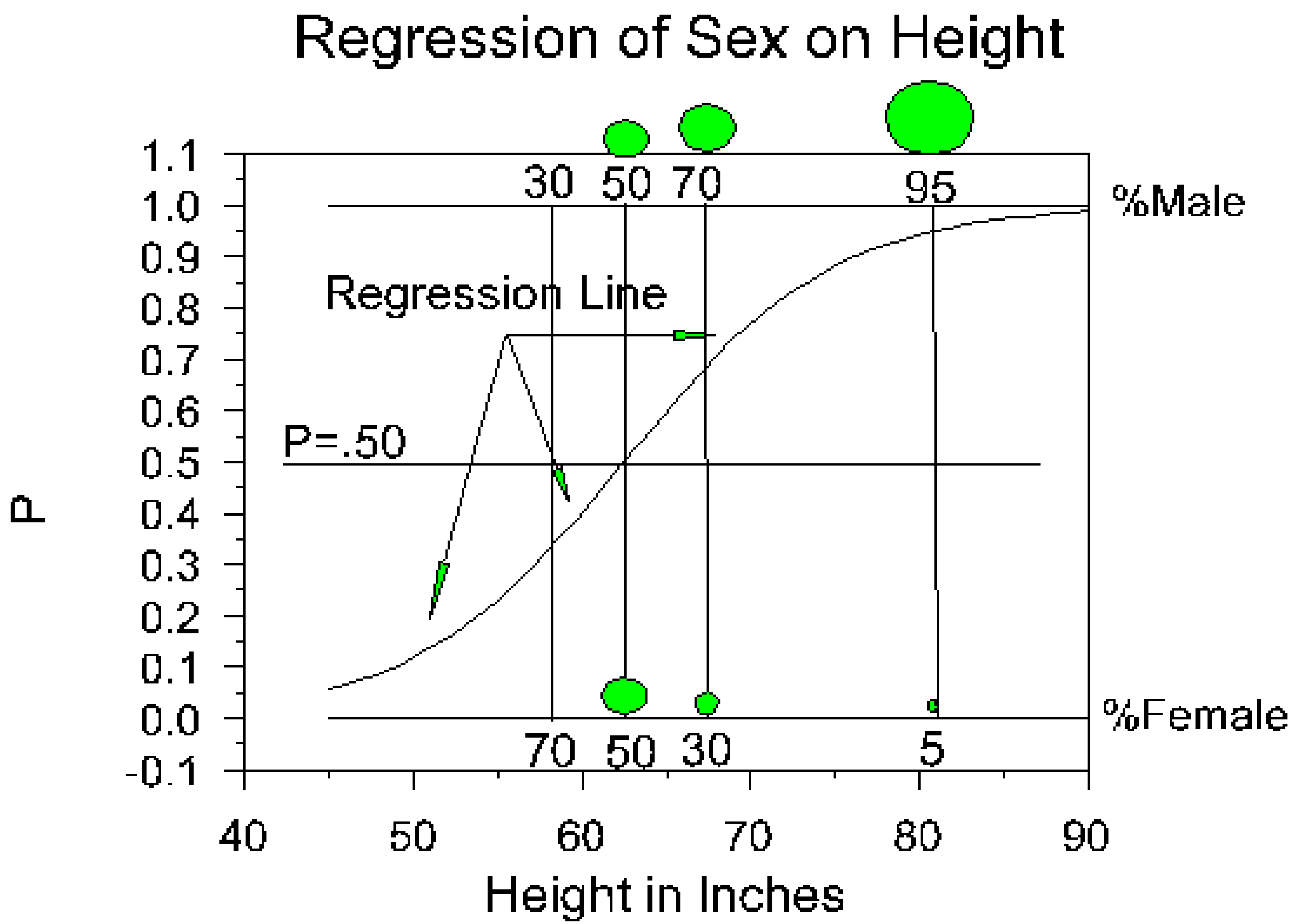
- In other words
  - $Y|\mathrm{X}, \beta \sim Bernoulli(\sigma\left(X^T\beta\right))$

The 1-d Situation

- $P(Y = 1|x, \beta_0, \beta_1) = \sigma(\beta_1 x + \beta_0)$



Example



Q: What is β_0 here?

Parameter Estimation

- MLE estimation
  - Given a dataset $D, with\ n\ data\ points$
  - For a single data object with attributes $\boldsymbol{x}_i$, class label y_i
    - Let $p_i = p(Y = 1|\boldsymbol{x}_i, \beta), the\ prob.of\ i\ in\ class\ 1$
    - The probability of observing y_i would be
      - If $y_i = 1, then\ \ p_i$
      - If $y_i = 0, then\ 1 - p_i$
      - Combing the two cases: $p_i^{y_i}(1 - p_i)^{1-y_i}$

$$L = \prod_i p_i^{y_i}(1 - p_i)^{1-y_i} = \prod_i \left(\frac{\exp\{\boldsymbol{x}_i^T\beta\}}{1+\exp\{\boldsymbol{x}_i^T\beta\}}\right)^{y_i} \left(\frac{1}{1+\exp\{\boldsymbol{x}_i^T\beta\}}\right)^{1-y_i}$$

Optimization

- Equivalent to maximize log likelihood
  - $L = \sum_i (y_i \boldsymbol{x}_i^T \beta - \log(1 + \exp\{\boldsymbol{x}_i^T \beta\}))$
- Gradient ascent update:
  - $$\beta^{new} = \beta^{old} + \eta \frac{\partial L(\beta)}{\partial \beta}$$

Step size

- Newton-Raphson update
  - $$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial L(\beta)}{\partial \beta}$$
  - where derivatives are evaluated at β^{old}

First Derivative

$$\frac{\partial L(\beta)}{\beta_{1j}} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} \frac{x_{ij} e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}$$

$p(x_i; \beta)$

$$= \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} p(x;\beta) x_{ij}$$

$$= \sum_{i=1}^{N} x_{ij}(y_i - p(x_i;\beta))$$

j = 0, 1, …, p

Second Derivative

- It is a (p+1) by (p+1) matrix, Hessian Matrix, with jth row and nth column as

$$
\begin{aligned}
&\frac{\partial L(\beta)}{\partial \beta_{1j} \partial \beta_{1n}} \\
= \quad & -\sum_{i=1}^{N} \frac{(1+e^{\beta^T x_i}) e^{\beta^T x_i} x_{ij} x_{in} - (e^{\beta^T x_i})^2 x_{ij} x_{in}}{(1+e^{\beta^T x_i})^2} \\
= \quad & -\sum_{i=1}^{N} x_{ij} x_{in} p(x_i;\beta) - x_{ij} x_{in} p(x_i;\beta)^2 \\
= \quad & -\sum_{i=1}^{N} x_{ij} x_{in} p(x_i;\beta)(1-p(x_i;\beta)) \, .
\end{aligned}
$$

An Alternative View of the Objective Function

- Cross entropy loss
  - Measure the difference from the predicted distribution (p) to the ground truth distribution (q)
    - Cross entropy from $q\ to\ p$: $H(q,p) = -\sum_k q_k \log(p_k)$
  - In the classification setting
    - $q_0 = 1\ and\ q_1 = 0, if\ y = 0; q_0 = 0\ and\ q_1 = 1, if\ y = 1$
    - $p_0 = \frac{1}{1+\exp\{\boldsymbol{x}^T\beta\}}$ and $p_1 = \frac{\exp\{\boldsymbol{x}^T\beta\}}{1+\exp\{\boldsymbol{x}^T\beta\}}$

An Alternative View of the Objective Function (Cont.)

- If $y = 0$
  - $H(q, p) = \log(1 + exp\{\boldsymbol{x}^T\beta\})$
- If $y = 1$
  - $H(q, p) = -\boldsymbol{x}^T\beta + \log(1 + exp\{\boldsymbol{x}^T\beta\})$

- Putting together
  - $H(q, p) = -y\boldsymbol{x}^T\beta + \log(1 + exp\{\boldsymbol{x}^T\beta\})$

- The goal: minimize the mean cross entropy loss over all the data points

What about Multiclass Classification?

- It is easy to handle under logistic regression, say M classes, using softmax function
  - $p_j = P(Y = j|X) = \frac{\exp\{X^T \beta_j\}}{1+\sum_{m=1}^{M-1} \exp\{X^T \beta_m\}}$, for j $= 1, \ldots, M-1$
  - $p_M = P(Y = M|X) = \frac{1}{1+\sum_{m=1}^{M-1} \exp\{X^T \beta_m\}}$
- Loss function
  - Cross entropy loss from observed class distribution (e.g., (0,0,1,0,0)) to p

Vector Data: Logistic Regression

- Classification: Basic Concepts
- Logistic Regression Model
- Generalized Linear Model*
- Summary

Recall Linear Regression and Logistic Regression

- Linear Regression
  - $y|\boldsymbol{x}, \beta \sim N(\boldsymbol{x}^T\beta, \sigma^2)$
- Logistic Regression
  - $y|\boldsymbol{x}, \beta \sim Bernoulli(\sigma(\boldsymbol{x}^T\beta))$
- How about other distributions?
  - Yes, as long as they belong to exponential family

Exponential Family

- Canonical Form
  - $p(y;\eta) = b(y)\exp\left(\eta^T T(y) - a(\eta)\right)$
  - η: natural parameter
  - $T(y)$: sufficient statistic
  - $a(\eta)$: log partition function for normalization
  - $b(y)$: function that only dependent on y

Examples of Exponential Family

- Many:

$$p(y;\eta) = b(y)\exp\left(\eta^T T(y) - a(\eta)\right)$$

  - Gaussian, Bernoulli, Poisson, beta, Dirichlet, categorical, …

- For Gaussian (not interested in σ)

$$\begin{aligned} p(y;\mu) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\cdot\exp\left(\mu y - \frac{1}{2}\mu^2\right) \end{aligned}$$

$$\begin{aligned} \eta &= \mu \\ T(y) &= y \\ a(\eta) &= \mu^2/2 \\ &= \eta^2/2 \\ b(y) &= (1/\sqrt{2\pi})\exp(-y^2/2) \end{aligned}$$

- For Bernoulli

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\underbrace{\boxed{\log\left(\frac{\phi}{1-\phi}\right)}}_{\eta}\right)y + \log(1-\phi)\right) \end{aligned}$$

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\log(1-\phi) \\ &= \log(1+e^\eta) \\ b(y) &= 1 \end{aligned}$$

Recipe of GLMs

- Determines a distribution for y
  - E.g., Gaussian, Bernoulli, Poisson
- Form the linear predictor for η
  - $\eta = \boldsymbol{x}^T \beta$
- Determines a link function: $\mu = g^{-1}(\eta)$
  - Connects the linear predictor to the mean of the distribution
  - E.g., $\mu = \eta$ for Gaussian, $\mu = \sigma(\eta)$ for Bernoulli, $\mu = exp(\eta)$ for Poisson

Vector Data: Logistic Regression

- Classification: Basic Concepts
- Logistic Regression Model
- Generalized Linear Model*
- Summary

Summary

- What is classification
  - Supervised learning vs. unsupervised learning, classification vs. prediction
- Logistic regression
  - Sigmoid function, multiclass classification
- Generalized linear model*
  - Exponential family, link function