PROBABILISTIC MODELS FOR STRUCTURED DATA

02: Naïve Bayes and Logistic Regression

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

January 10, 2019

Content

- Probabilistic Models for I.I.D. Data
- Naïve Bayes
- Logistic Regression
- Generative Models and Discriminative Models
- Summary

I.I.D. Data

- Data: $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$
  - A data point $(\boldsymbol{x}_i, y_i)$ contains a feature vector and a label
  - n: number of data points
- Assume data points are independent and identically distributed (i.i.d.)
  - Model $p(D|\theta)$ under I.I.D. assumption
    - $p(D|\theta) = \prod_i p(\boldsymbol{x}_i, y_i|\theta)$ (if modeling joint distribution)
    - $p(D|\theta) = \prod_i p(y_i|\boldsymbol{x}_i, \theta)$ (if modeling conditional distribution, conditional i.i.d.)
  - Inference under I.I.D. assumption
    - Inference can be made for individual data points indepedently

Content

- Probabilistic Models for I.I.D. Data
- Naïve Bayes
- Logistic Regression
- Generative Models and Discriminative Models
- Summary

Naïve Bayes for Text

- Text Data
- Revisit of Multinomial Distribution
- Multinomial Naïve Bayes

Text Data

- Word/term
- Document
  - A sequence of words
- Corpus
  - A collection of documents

Text Classification Applications

- Spam detection

From: airak@medicana.com.tr
Subject: Loan Offer
Do you need a personal or business loan urgent that can be process within 2 to 3 working days? Have you been frustrated so many times by your banks and other loan firm and you don't know what to do? Here comes the Good news Deutsche Bank Financial Business and Home Loan is here to offer you any kind of loan you need at an affordable interest rate of 3% If you are interested let us know.

- Sentiment analysis

The Lion King, complete with jaunty songs by Elton John and Tim Rice, is undeniably and fully worthy of its glorious Disney heritage. It is a gorgeous triumph -- one lion in which the studio can take justified pride.

Between traumas, the movie serves up soothingly banal musical numbers (composed by Elton John and Tim Rice) and silly, rambunctious comedy.

July 31, 2013 | Full Review...

Represent A Document

- A document d is represented by a sequence of words selected from a vocabulary
  - $\boldsymbol{w}_d = \left(w_{d1}, w_{d2}, \ldots, w_{dN_d}\right)$, where w_{di} is the id of i-th word in document d and N_d is the length of document d
- A bag-of-words representation
  - $\boldsymbol{x}_d = (x_{d1}, x_{d2}, \ldots, x_{dN})$, where x_{dn} is the number of words for nth word in the vocabulary
  - $x_{dn} = \sum_i 1(w_{di} == n)$

Example

c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

$\boldsymbol{x}_d$

Naïve Bayes for Text

- Text Data
- Revisit of Multinomial Distribution
- Multinomial Naïve Bayes

Bernoulli and Categorical Distribution

- Bernoulli distribution
  - Discrete distribution that takes two values $\{0,1\}$
    - $P(X = 1) = p$ and $P(X = 0) = 1 - p$
    - E.g., toss a coin with head and tail

- Categorical distribution
  - Discrete distribution that takes more than two values, i.e., $x \in \{1, \dots, K\}$
    - Also called generalized Bernoulli distribution, multinoulli distribution
    - $P(X = k) = p_k \; and \; \sum_k p_k = 1$
    - E.g., get 1-6 from a dice with 1/6

Binomial and Multinomial Distribution

- Binomial distribution
  - Number of successes (i.e., total number of 1's) by repeating n trials of independent Bernoulli distribution with probability p
    - $x: number\ of\ successes$
    - $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
- Multinomial distribution (multivariate random variable)
  - Repeat n trials of independent categorical distribution
    - Let x_k be the number of times value k has been observed, note $\sum_k x_k = n$
    - $P(X_1 = x_1, X_2 = x_2, \dots, X_K = x_K) = \frac{n!}{x_1! x_2! \dots x_K!} \prod_k p_k^{x_k}$

Naïve Bayes for Text

- Text Data
- Revisit of Multinomial Distribution
- Multinomial Naïve Bayes

Bayes' Theorem: Basics

- Bayes' Theorem: $P(h|\mathbf{X}) = \frac{P(\mathbf{X}|h)P(h)}{P(\mathbf{X})}$
  - Let **X** be a data sample ("*evidence*")
  - Let h be a *hypothesis* that X belongs to class C
  - P(h) (*prior probability*): the probability of hypothesis h
    - E.g., the probability of "spam" class
  - P(**X**|h) (*likelihood*): the probability of observing the sample **X**, given that the hypothesis holds
    - E.g., the probability of an email given it's a spam
  - P(**X**): marginal probability that sample data is observed
    - $P(X) = \sum_h P(X|h)\,P(h)$
  - P(h|**X**), (i.e., *posterior probability*): the probability that the hypothesis holds given the observed data sample **X**

Classification: Choosing Hypotheses

- *Maximum a posteriori* (maximize the posterior):
  - Useful observation: it does not depend on the denominator P(X)

$$h_{MAP} = \arg\max_{h \in H} P(h \mid X) = \arg\max_{h \in H} P(X \mid h) P(h)$$

Classification by Maximum A Posteriori

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an p-D attribute vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$

- Suppose there are *m* classes $y \in \{1, 2, \ldots, m\}$

- Classification is to derive the maximum posteriori, i.e., the maximal $P(y=j|\mathbf{x})$

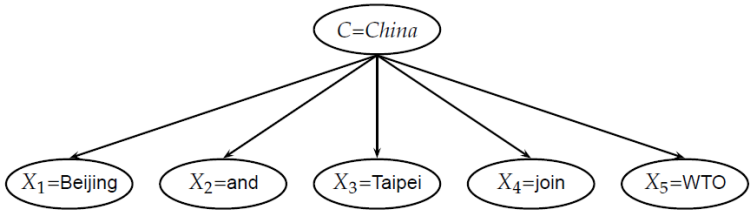
- This can be derived from Bayes' theorem

$$p(y = j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y = j)p(y = j)}{p(\boldsymbol{x})}$$

- Since $p(\mathbf{x})$ is constant for all classes, only $p(\boldsymbol{x}|y)p(y)$ needs to be maximized

Now Come to Text Setting: Modeling

- A document is represented as
  - $\boldsymbol{w}_d = (w_{d1}, w_{d2}, \dots, w_{dN_d})$
  - w_{di} is the i-th word of d and N_d is the length of document d
- Model $p(\boldsymbol{w}_d|y)$ for class y
  - Each word in the sequence w_{di} is sampled from multinoulli distribution with parameter vector $\boldsymbol{\beta}_y = (\beta_{y1}, \beta_{y2}, \dots, \beta_{yN})$ independently
    - $p(w_{di}|y) = \beta_{yw_{di}}$ and $p(\boldsymbol{w}_d|y) = \prod_i \beta_{yw_{di}} = \prod_n \beta_{yn}^{x_{dn}}$
    - Where x_{dn} is the number of words for nth word in the vocabulary
- Model $p(y = j)$
  - Follow categorical distribution with parameter vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_m)$, i.e.,
    - $p(y = j) = \pi_j$



Classification Process Assuming Parameters are Given: Inference

- Find y that maximizes $p(y|\boldsymbol{x}_d)$, which is equivalently to maximize

$$y^* = \underset{y}{argmax}\ p(\boldsymbol{x}_d, y)$$

$$= argmax_y\ p(\boldsymbol{x}_d|y)p(y)$$

$$= argmax_y \prod_n \beta_{yn}^{x_{dn}} \times \pi_y$$

$$= argmax_y \prod_n \beta_{yn}^{x_{dn}} \times \pi_y$$

$$= argmax_y \sum_n x_{dn} log\beta_{yn} + log\pi_y$$

Parameter Estimation via MLE: Learning

- Given a corpus and labels for each document
  - $D = \{(\boldsymbol{x}_d, y_d)\}$
  - Find the MLE estimators for $\Theta = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_m, \boldsymbol{\pi})$
- The log likelihood function for the training dataset

$$logL(\Theta) = log \prod_d p(\boldsymbol{x}_d, y_d|\Theta) = \sum_d log\, p(\boldsymbol{x}_d, y_d|\Theta)$$

$$= \sum_d log\, p(\boldsymbol{x}_d|y_d)p(y_d) = \sum_d (x_{dn} log\beta_{y_d n} + log\pi_{y_d})$$

- The optimization problem

$$\max_{\Theta} \log L\,(\Theta)$$

$$s.t.$$

$$\pi_j \geq 0\ and\ \sum_j \pi_j = 1$$

$$\beta_{jn} \geq 0\ and\ \sum_n \beta_{jn} = 1\ for\ all\ j$$

Solve the Optimization Problem

- Use the Lagrange multiplier method
- Solution
  - $\hat{\beta}_{jn} = \frac{\sum_{d:y_d==j} x_{dn}}{\sum_{d:y_d==j} \sum_{n'} x_{dn'}}$
    - $\sum_{d:y_d=j} x_{dn}$: total count of word n in class j
    - $\sum_{d:y_d=j} \sum_{n'} x_{dn'}$: total count of words in class j
  - $\hat{\pi}_j = \frac{\sum_d 1(y_d==j)}{|D|}$
    - $1(y_d = j)$ is the indicator function, which equals to 1 if $y_d = j$ holds
    - |D|: total number of documents

Smoothing

- What if some word n does not appear in some class j in training dataset?
  - $\hat{\beta}_{jn} = \frac{\sum_{d:y_d=j} x_{dn}}{\sum_{d:y_d=j} \sum_{n'} x_{dn'}} = 0$
  - $\Rightarrow p(\boldsymbol{x}_d | y = j) \propto \prod_n \beta_{jn}^{x_{dn}} = 0$
  - But other words may have a strong indication the document belongs to class j
- Solution: add-1 smoothing or Laplace smoothing
  - $\hat{\beta}_{jn} = \frac{\sum_{d:y_d=j} x_{dn}+1}{\sum_{d:y_d=j} \sum_{n'} x_{dn'}+N}$
  - N: $total\ number\ of\ words\ in\ the\ vocabulary$
  - Check: $\sum_n \hat{\beta}_{jn} = 1$?

Example

- Data:

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

- Vocabulary:

Index	1	2	3	4	5	6
Word	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan

- Learned parameters (with smoothing):

$$\hat{\beta}_{c1} = \frac{5+1}{8+6} = \frac{3}{7}$$
$$\hat{\beta}_{c2} = \frac{1+1}{8+6} = \frac{1}{7}$$
$$\hat{\beta}_{c3} = \frac{1+1}{8+6} = \frac{1}{7}$$
$$\hat{\beta}_{c4} = \frac{1+1}{8+6} = \frac{1}{7}$$
$$\hat{\beta}_{c5} = \frac{0+1}{8+6} = \frac{1}{14}$$
$$\hat{\beta}_{c6} = \frac{0+1}{8+6} = \frac{1}{14}$$

$$\hat{\beta}_{j1} = \frac{1+1}{3+6} = \frac{2}{9}$$
$$\hat{\beta}_{j2} = \frac{0+1}{3+6} = \frac{1}{9}$$
$$\hat{\beta}_{j3} = \frac{0+1}{3+6} = \frac{1}{9}$$
$$\hat{\beta}_{j4} = \frac{0+1}{3+6} = \frac{1}{9}$$
$$\hat{\beta}_{j5} = \frac{1+1}{3+6} = \frac{2}{9}$$
$$\hat{\beta}_{j6} = \frac{1+1}{3+6} = \frac{2}{9}$$

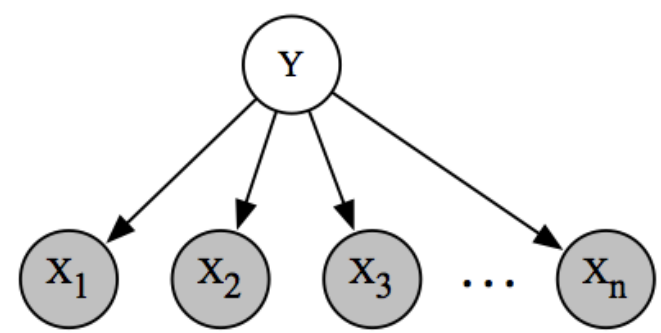
$$\hat{\pi}_c = \frac{3}{4}$$
$$\hat{\pi}_j = \frac{1}{4}$$

Example (Continued)

- Classification stage
  - For the test document d=5, compute
  - $p(y = c|\boldsymbol{x}_5) \propto p(y = c) \times \prod_n \beta_{cn}^{x_{5n}} = \frac{3}{4} \times \left(\frac{3}{7}\right)^3 \times \left(\frac{1}{14}\right) \times \left(\frac{1}{14}\right) \approx 0.0003$
  - $p(y = j|\boldsymbol{x}_5) \propto p(y = j) \times \prod_n \beta_{jn}^{x_{5n}} = \frac{1}{4} \times \left(\frac{2}{9}\right)^3 \times \left(\frac{2}{9}\right) \times \left(\frac{2}{9}\right) \approx 0.0001$
  - Conclusion: $\boldsymbol{x}_5$ should be classified into c class

A More General Naïve Bayes Framework

- Let D be a training set of tuples and their class labels, and each tuple is represented by an p-D attribute vector **x** = (x_1, x_2, …, x_p)
- Suppose there are *m* classes y∈{1, 2, …, m}
- Goal: Find $\mathrm{y} = \arg\max_{y} p(y|\boldsymbol{x}) = p(y, \boldsymbol{x})/p(\boldsymbol{x}) \propto p(\boldsymbol{x}|y)p(y)$
- A simplified assumption: attributes are conditionally independent given the class (class conditional independency):
  - $p(\boldsymbol{x}|y) = \prod_k p(x_k|y)$
  - $p(x_k|y)$ can follow any distribution,
    - e.g., Gaussian, Bernoulli, categorical, …



Content

- Probabilistic Models for I.I.D. Data
- Naïve Bayes
- Logistic Regression
- Generative Models and Discriminative Models
- Summary

Linear Regression VS. Logistic Regression

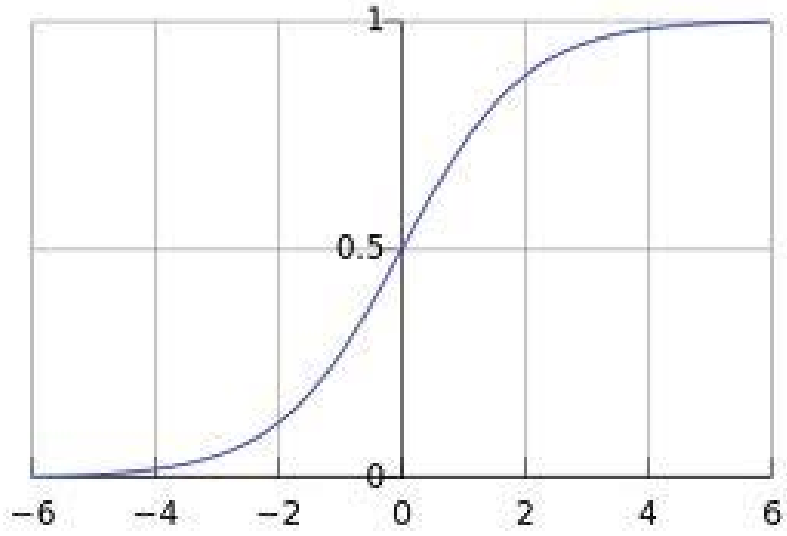
- Linear Regression (prediction)
  - Y: $continuous\ value\ (-\infty, +\infty)$
    - $\mathrm{y} = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$
    - $\mathrm{y}|\boldsymbol{x}, \beta \sim N(\boldsymbol{x}^T\beta, \sigma^2)$

- Logistic Regression (classification)
  - Y: $discrete\ value\ from\ m\ classes$
    - $P(Y = j|\boldsymbol{x}, \boldsymbol{\beta}) \in [0,1]\ and\ \sum_j P(Y = j|\boldsymbol{x}, \boldsymbol{\beta}) = 1$

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

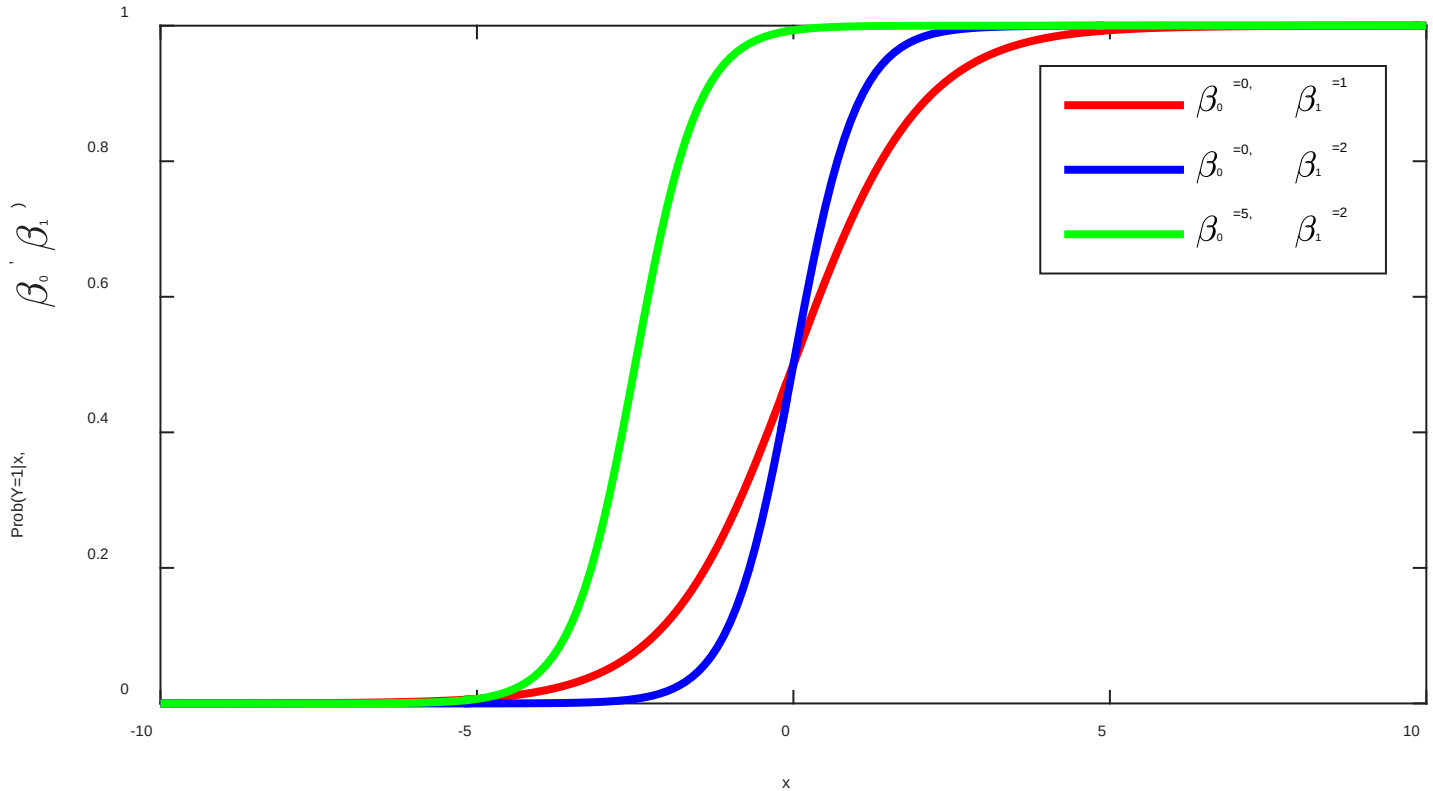
- $P(Y = 1|\boldsymbol{x}, \beta) = \sigma\left(\boldsymbol{x}^T\beta\right) = \frac{1}{1+\exp\{-\boldsymbol{x}^T\beta\}} = \frac{\exp\{\boldsymbol{x}^T\beta\}}{1+\exp\{\boldsymbol{x}^T\beta\}}$
- $P(Y = 0|\boldsymbol{x}, \beta) = 1 - \sigma\left(\boldsymbol{x}^T\beta\right) = \frac{\exp\{-\boldsymbol{x}^T\beta\}}{1+\exp\{-\boldsymbol{x}^T\beta\}}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

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

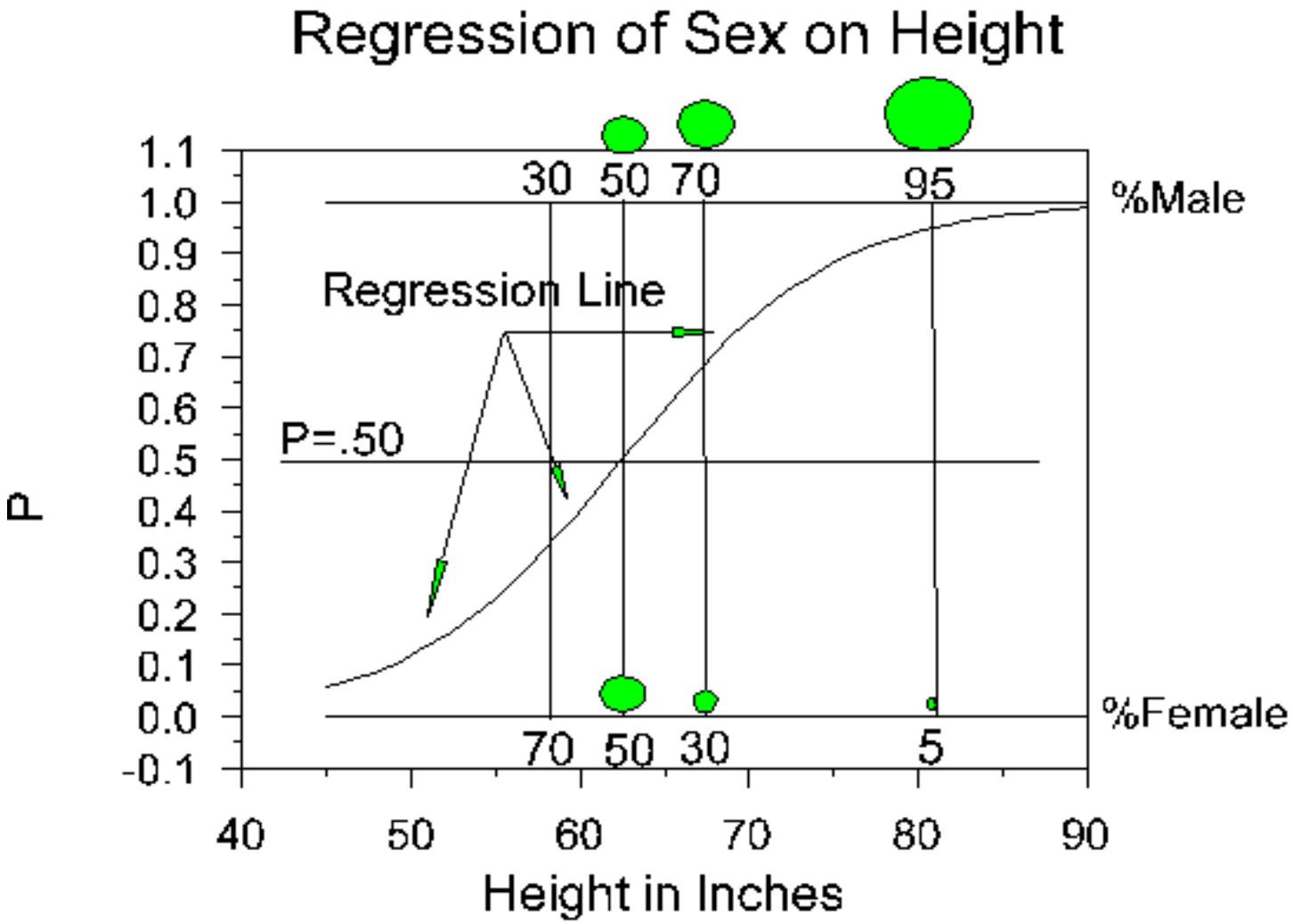
- In other words
  - $\mathrm{y}|\boldsymbol{x}, \beta \sim Bernoulli(\sigma\left(\boldsymbol{x}^T\beta\right))$

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?

Classification Assuming Parameters are Given: Inference

- $If\ P(Y = 1|\boldsymbol{x}, \beta) = \sigma\left(\boldsymbol{x}^T\beta\right) > 0.5$
  - Class 1
- Otherwise
  - Class 0

Parameter Estimation: Learning

- 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 $p_i = p(y_i = 1|\boldsymbol{x}_i, \beta), the\ prob. of\ i\ in\ class\ 1$
    - 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: $p_i^{y_i}(1 - p_i)^{1-y_i}$

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

Optimization

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

Step size

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

First Derivative

- It is a (p+1) vector, with jth element as

$$\frac{\partial log L(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} \frac{x_{ij} e^{\beta^T \mathbf{x}_i}}{1 + e^{\beta^T \mathbf{x}_i}}$$

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

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

$p_i(\beta)$

For j = 0, 1, …, p

Second Derivative

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

$$\frac{\partial log L(\beta)}{\partial \beta_j \partial \beta_n} = -\sum_{i=1}^{N} \frac{(1+e^{\beta^T \mathbf{x}_i})e^{\beta^T \mathbf{x}_i} x_{ij} x_{in} - (1+e^{\beta^T \mathbf{x}_i})^2 x_{i}}{(1+e^{\beta^T \mathbf{x}_i})^2}$$

$$= -\sum_{i=1}^{N} x_{ij} x_{in} p_i(\beta) - \sum_{i=1}^{N} x_{ij} x_{in} (p_i(\beta))^2$$

$$= -\sum_{i=1}^{N} x_{ij} x_{in} p_i(\beta)(1 - p_i(\beta))$$

What about Multiclass Classification?

- It is easy to handle under logistic regression, say M classes
  - $P(Y = j|x) = \frac{\exp\{x^T\beta_j\}}{1+\sum_{m=1}^{M-1}\exp\{x^T\beta_m\}}$, for j = $1, \dots, M-1$
  - $P(Y = M|x) = \frac{1}{1+\sum_{m=1}^{M-1}\exp\{x^T\beta_m\}}$

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:

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

  - 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

Content

- Probabilistic Models for I.I.D. Data
- Naïve Bayes
- Logistic Regression
- Generative Models and Discriminative Models
- Summary

Generative Models vs. Discriminative Models

- Generative model
  - $model\ joint\ probability\ p(\boldsymbol{x}, y)$
  - E.g., naïve Bayes
- Discriminative model
  - $model\ conditional\ probability\ p(y|\boldsymbol{x})$
  - E.g., logistic regression

Which One is Better?

- Consider $p(\boldsymbol{x}, y) = p(y|\boldsymbol{x}) \times p(\boldsymbol{x})$
  - Generative models require additional model of marginal distribution $p(\boldsymbol{x})$
    - Need more data to learn $p(\boldsymbol{x})$
    - Distribution assumption of $p(\boldsymbol{x})$ might be incorrect
  - In practice, discriminative models work very well

https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf

Content

- Probabilistic Models for I.I.D. Data
- Naïve Bayes
- Logistic Regression
- Generative Models and Discriminative Models
- Summary

Summary

- Probabilistic Models for I.I.D. Data
  - I.I.D. assumption enables joint distribution of data as a product of probability of single data points
- Naïve Bayes
  - Assuming independence among features
- Logistic Regression
  - Assuming conditional distribution follows Bernoulli distribution
- Generative Models and Discriminative Models
  - Modeling joint distribution vs. conditional distribution

References

- http://pages.cs.wisc.edu/~jerryzhu/cs769/nb.pdf
- http://cs229.stanford.edu/notes/cs229-notes1.pdf
- https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf

More about Lagrangian

- Objective with equality constraints

$$\min_{w} f(w)$$

$$s.t.$$

$$h_i(w) = 0, for\ i = 1,2, \dots, l$$

- Lagrangian:
  - $L(w, \boldsymbol{\alpha}) = f(w) + \sum_i \alpha_i h_i(w)$
    - α_i: Lagrangian multipliers
- Solution: setting the derivatives of Lagrangian to be 0
  - $\frac{\partial L}{\partial w} = 0\ and\ \frac{\partial L}{\partial \alpha_i} = 0$ for every i

Generalized Lagrangian

- Objective with both equality and inequality constraints

$$\min_{w} f(w)$$
$$s.t.$$
$$h_i(w) = 0, for\ i = 1,2,\dots,l$$
$$g_j(w) \leq 0, for\ j = 1,2,\dots,k$$

- Lagrangian
  - $L(w, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(w) + \sum_i \alpha_i h_i(w) + \sum_j \beta_j g_j(w)$
    - α_i: Lagrangian multipliers
    - $\beta_j \geq 0$: Lagrangian multipliers

Why It Works

- Consider function

$$\theta_p(w) = \max_{\alpha,\beta:\beta_j \geq 0} L(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- $\theta_p(w) = \begin{cases} f(w), & if\ w\ satisfies\ all\ constraints \\ \infty, & if\ w\ doesn't\ satisfy\ constraints \end{cases}$
- Therefore, minimize $f(w)$ with constraints is equivalent to minimize $\theta_p(w)$

Lagrange Duality

- The primal problem

$$p^* = \min_{w} \max_{\alpha,\beta:\beta_j \geq 0} L(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- The dual problem

$$d^* = \max_{\alpha,\beta:\beta_j \geq 0} \min_{w} L(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- According to max-min inequality

$$p^* \leq d^*$$

  - When does equation hold?

Primal = Dual

- $p^* = d^*$, under some proper condition (Slater conditions)
  - f, g_j convex, h_i affine
  - Exists w, such that all $g_j(\mathrm{w}) < 0$
- (w^*, α^*, β^*) need to satisfy KKT conditions
  - $\frac{\partial L}{\partial w} = 0$
  - $\beta_j g_j(w) = 0$
  - $h_i(w) = 0, g_j(w) \leq 0, \beta_j \geq 0$

https://cs.stanford.edu/people/davidknowles/lagrangian_duality.pdf