Statistical Pattern Recognition
Lecture3
Logistic Regression

Dr Zohreh Azimifar

School of Electrical and Computer Engineering

Shiraz University

Fall2014

Table of contents

Introduction

- Classification: given class labels are discrete values.
- Class: a category of patterns (c classes).
- Goal: mapping from feature vectors to class labels.
- Probabilistic view of classification:

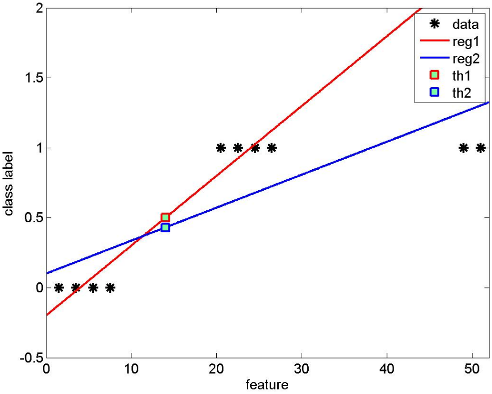
$$y^* = argmax_y \; P(y|\mathbf{X})$$

  y^* is the class label which maximizes probability of labels given the pattern $\mathbf{X}$.

- In probabilistic classification we train for this probability density function.

Introduction

- Linear Regression for classification?



Discrete Labels

- In a 2-class problem ($y \in \{0, 1\}$), we use probability density to predict class label:

$$P(y = 1|\mathbf{X}; \boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\mathbf{X})$$
$$P(y = 0|\mathbf{X}; \boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(\mathbf{X})$$

  therefore:

$$P(y|\mathbf{X}; \boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\mathbf{X})^y (1 - h_{\boldsymbol{\theta}}(\mathbf{X}))^{1-y}$$

- Here $h_{\boldsymbol{\theta}}(\mathbf{X})$ is the hypothesis model showing probability for pattern $\mathbf{X}$ to be of class 1.

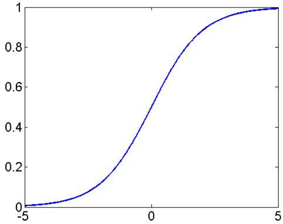
Sigmoid Function

- Here $h_{\theta}(\mathbf{X})$ is the hypothesis model showing probability for pattern $\mathbf{X}$ to be of class 1

$$h_{\theta}(\mathbf{X}) = g(\theta^T\mathbf{X}) = \frac{1}{1+e^{-\theta^T\mathbf{X}}}$$

$$g(z) = \frac{1}{1+e^{-z}}$$

- Function $g(.)$ is called Sigmoid or logistic function:



Maximum Likelihood Estimate

- Model parameters are determined by maximum likelihood (ML) estimate:

$$\begin{aligned} L(\boldsymbol{\theta}) &= P(\mathbf{y}|X;\boldsymbol{\theta}) \\ &= \prod_{j=1}^{m} P(y^{(j)}|\mathbf{X}^{(j)};\boldsymbol{\theta}) \\ &= \prod_{j=1}^{m} h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})^{y^{(j)}}(1-h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)}))^{(1-y^{(j)})} \end{aligned}$$

- Samples are assumed independently and identically distributed(*i.i.d.*).

Log Likelihood

- In this case maximizing the log likelihood is mathematically easier

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= logL(\boldsymbol{\theta}) \\
&= \sum_{j=1}^{m} \{ y^{(j)} log(h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})) + (1 - y^{(j)}) log(1 - h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})) \}
\end{aligned}
$$

- $l(\boldsymbol{\theta})$ is to be maximized. Use gradient ascent

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

Gradient Ascent Method

$$\frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta}) = \sum_{j=1}^{m} \{ y^{(j)} \frac{h'_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})}{h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})} + (1 - y^{(j)}) \frac{-h'_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})}{1 - h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})} \}$$

where:

$$\frac{\partial}{\partial \theta_i} h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)}) = h'_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})$$

$$= \frac{x_i^{(j)} e^{-\boldsymbol{\theta}^T \mathbf{X}}}{(1 + e^{-\boldsymbol{\theta}^T \mathbf{X}})^2} = x_i^{(j)} h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})(1 - h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)}))$$

substitute back:

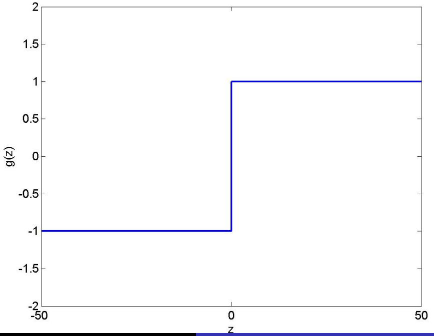
$$\frac{\partial}{\partial \theta_i} \; l(\boldsymbol{\theta}) = \sum_{j=1}^{m} (y^{(j)} - h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})) x_i^{(j)}$$

$$\theta_i = \theta_i + \alpha \sum_{j=1}^{m} (y^{(j)} - h_{\boldsymbol{\theta}}(\mathbf{X}^{(j)})) x_i^{(j)}$$

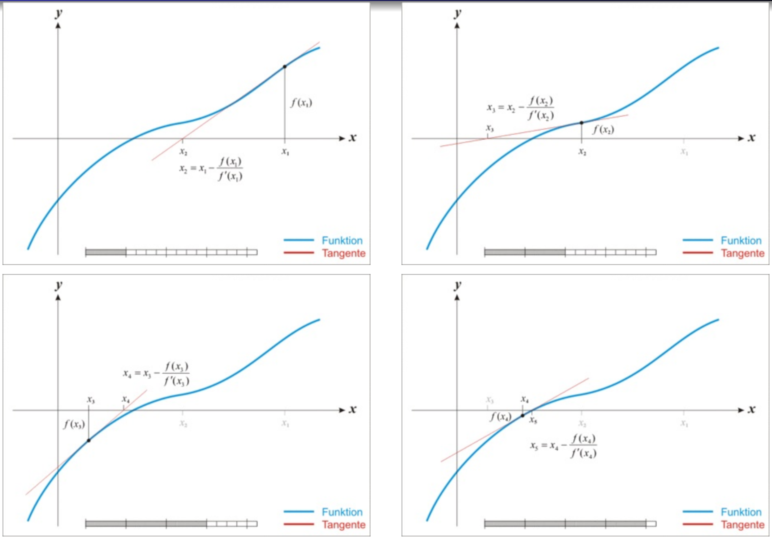
Linearity of The Model

- Where does linearity of logistic regression come from?
- Any other choice for the hypothesis function?

$$g(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{otherwise} \end{cases}$$



Parameter Learning Using Newton's Method



Parameter Learning Using Newton's Method

- Newton's method finds root of an arbitrary function $f(\theta)$, iteratively:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})}$$

- Recall, the objective is to determine parameters of likelihood function $l(\theta)$:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{l'(\theta^{(t)})}{l''(\theta^{(t)})}$$

  to be generalized for all parameters $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - H^{-1}\nabla_{\boldsymbol{\theta}} l$$

  where H is called Hessian matrix:

$$H_{ij} = \frac{\partial^2 l}{\partial\theta_i \partial\theta_j}$$

Exponential Distribution Family

- Can we extend linear regression and logistic regression to more general models?
- Objective: determine general models for $P(y|\mathbf{X};\boldsymbol{\theta})$.
- Define exponential distribution family:

$$P(y;\eta) = b(y)\ exp(\eta^T T(y) - a(\eta))$$

where η is Natural Parameter T is Sufficient Statistic $T(y) = y$

- Different selection of $\{a,\ b,\ T,\ \eta\}$ yields different distributions:
  - Bernoulli Distribution
  - Gaussian Distribution

Bernoulli Distribution

$$P(y=1;\phi) = 1 - P(y=0;\phi) = \phi$$

- Special case of the general model?

$$\begin{aligned} P(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= exp(log(\phi^y(1-\phi)^{1-y})) \\ &= exp(ylog\phi + (1-y)log(1-\phi)) \\ &= \underbrace{1}_{b(y)} \times\ exp(\underbrace{(log\frac{\phi}{1-\phi})}_{\eta}\underbrace{y}_{T(y)} + \underbrace{log(1-\phi)}_{-a(\eta)}) \end{aligned}$$

therefore:

$$\eta = log\frac{\phi}{1-\phi} \Longrightarrow \phi = \frac{1}{1+e^{-\eta}}$$

$$a(\eta) = -log(1-\phi) = log(1+e^{\eta})$$

Gaussian Distribution

$$N(\mu, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} exp(\frac{-1}{2}(y-\mu)^2)$$

- Special case of the general model?

$$N(\mu, \sigma^2) = \underbrace{\frac{1}{\sqrt{2\pi}} exp(\frac{-1}{2} y^2)}_{b(y)} \cdot exp(\underbrace{\mu}_{\eta} \underbrace{y}_{T(y)} - \frac{1}{2} \underbrace{\mu^2}_{a(\eta) = +\frac{1}{2}\mu^2 = +\frac{1}{2}\eta^2})$$

Conditions for General Linear Regression Models

Three conditions for a model $P(y|\mathbf{X};\boldsymbol{\theta})$ to be considered a linear model:

1. Model $P(y|\mathbf{X};\boldsymbol{\theta})$ be an exponential distribution with parameter η.
2. Given feature vector $\mathbf{X}$, predict $E[T(y)|\mathbf{X}]$.
   Here, we have $h_{\boldsymbol{\theta}}(\mathbf{X}) = E[T(y)|\mathbf{X}]$
3. $$\begin{cases} \eta = \boldsymbol{\theta}^T\mathbf{X} & \text{if } \eta \text{ is a real number} \\ \eta_i = \boldsymbol{\theta}_i^T\mathbf{X} & \text{if } \boldsymbol{\eta} \in \mathbb{R}^k \end{cases}$$

Logistics Regression and General Model

Three conditions for logistic regression to be considered a generalized linear model:

1. $$P(y|\mathbf{X};\boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\mathbf{X})^y (1 - h_{\boldsymbol{\theta}}(\mathbf{X}))^{1-y}$$
$$P(y=1;\phi) = 1 - P(y=0;\phi) = \phi$$

2. $$\begin{aligned} h_{\boldsymbol{\theta}}(\mathbf{X}) &= P(y=1|\mathbf{X};\boldsymbol{\theta}) \\ &= 1 \times P(y=1|\mathbf{X};\boldsymbol{\theta}) + 0 \times P(y=0|\mathbf{X};\boldsymbol{\theta}) \\ &= E[y|\mathbf{X}] \quad \text{(second condition satisfied)} \\ &= \phi = \frac{1}{1+e^{-\eta}} \\ &= \frac{1}{1+e^{-\boldsymbol{\theta}^T\mathbf{X}}} \quad \text{(third condition satisfied)} \end{aligned}$$

Linear Regression and General Model

Three conditions for linear regression to be considered a generalized linear model:

$$\begin{aligned} h_{\boldsymbol{\theta}}(\mathbf{X}) &= \boldsymbol{\theta}^T \mathbf{X} \\ &= \mu \\ &= E[y|\mathbf{X}] \\ &= \eta \end{aligned}$$

How do you describe the satisfiability here?

Multinomial Distribution

- Let us extend 2-class problem to c-class classification, *i.e.,* $y \in \{1, \ldots, c\}$.
- We assume the class multinomially distributed:

$$P(y = i|\mathbf{X}) = \phi_i \quad (i = 1, 2, \cdots, c)$$

  here, only determine $c - 1$ parameters, why?
- Try to fit a generalized linear model into this c-class problem.

General Linear Model

$$\mathbf{T}(1) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{T}(2) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, \mathbf{T}(c-1) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{T}(c) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \in \mathbb{R}^{c-1}$$

Define indicator function

$$\begin{cases} 1\{True\} = 1 \\ 1\{false\} = 0 \end{cases}$$

can represent value of the i^{th} element in T:

$$T(y)_i = 1\{y = i\}$$

General Linear Model

$$
\begin{aligned}
P(y|\mathbf{X};\phi) &= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\cdots\phi_c^{1\{y=c\}} \\
&= \phi_1^{T(y)_1}\phi_2^{T(y)_2}\cdots\phi_{c-1}^{T(y)_{c-1}}\phi_c^{1-\sum_{j=1}^{c-1}T(y)_j} \\
&= exp(T(y)_1 log(\phi_1) + T(y)_2 log(\phi_2) + \cdots \\
&\cdots + T(y)_{c-1} log(\phi_{c-1}) + (1 - \sum_{j=1}^{c-1} T(y)_j) log(\phi_c)) \\
&= exp(T(y)_1 log(\frac{\phi_1}{\phi_c}) + T(y)_2 log(\frac{\phi_2}{\phi_c}) + \cdots \\
&\cdots + T(y)_{c-1} log(\frac{\phi_{c-1}}{\phi_c}) + log(\phi_c)) \\
&= b(y) exp(\boldsymbol{\eta}^T T(y) - a(\boldsymbol{\eta}))
\end{aligned}
$$

General Linear Model

Aligning with the parameters of exponential distribution:

$$\boldsymbol{\eta} = \begin{bmatrix} log(\frac{\phi_1}{\phi_c}) \\ \vdots \\ log(\frac{\phi_{c-1}}{\phi_c}) \end{bmatrix} \in \mathbb{R}^{c-1}$$

$$a(\boldsymbol{\eta}) = -log(\phi_c)$$
$$b(y) = 1$$

therefore:

$$\phi_i = \frac{e^{\eta_i}}{1 + \sum_{j=1}^{c-1} e^{\eta_j}} \qquad (i = 1, \cdots, c-1)$$

General Linear Model

$$h_{\boldsymbol{\theta}}(\mathbf{X}) = E[T(y)|\mathbf{X};\boldsymbol{\theta}] = E\left[\begin{bmatrix} 1\{y=1\} \\ \vdots \\ 1\{y=c-1\} \end{bmatrix} \middle| \mathbf{X};\boldsymbol{\theta} \right] = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{c-1} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{e^{\eta_1}}{1+\sum_{j=1}^{c-1} e^{\eta_j}} \\ \vdots \\ \frac{e^{\eta_{c-1}}}{1+\sum_{j=1}^{c-1} e^{\eta_j}} \end{bmatrix} \qquad \text{(second condition)}$$

$$= \begin{bmatrix} \frac{e^{\boldsymbol{\theta}_1^T \mathbf{X}}}{1+\sum_{j=1}^{c-1} e^{\boldsymbol{\theta}_j^T \mathbf{X}}} \\ \vdots \\ \frac{e^{\boldsymbol{\theta}_{c-1}^T \mathbf{X}}}{1+\sum_{j=1}^{c-1} e^{\boldsymbol{\theta}_j^T \mathbf{X}}} \end{bmatrix} \qquad \text{(third condition } \eta_i = \boldsymbol{\theta}_i^T \mathbf{X}\text{)}$$

Likelihood Function for Softmax Regression

- $\phi_i = \frac{e^{\theta_i^T \mathbf{x}}}{1+\sum_{j=1}^{c-1} e^{\theta_j^T \mathbf{x}}}$ shows the probability for $\mathbf{X}$ being of class i.
- How many parameters does this model need to learn?
- Assuming *i.i.d.* sampels, the likelihood function is:

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{m} P(y^{(j)}|\mathbf{X}^{(j)};\boldsymbol{\theta})$$
$$= \prod_{j=1}^{m} \phi_1^{1\{y^{(j)}=1\}} \phi_2^{1\{y^{(j)}=2\}} \ldots \phi_c^{1\{y^{(j)}=c\}}$$

- Set derivative to zero and learn the parameters.

Summary

-
-