Logistic Regression

A brief tutorial

Jay

January 28, 2015

Outline

Outline

Basics

- Counterpart of Linear Regression
- Basic model for Classification
- Logistic: Uses the logit function
- Regression: Predicts the probability of an outcome

Outline

1. Definition
   - Basics
   - Parameters
   - Basic Model
   - Training
   - Testing
2. Motivation
   - Maximum Likelihood
   - Maximum Entropy
3. Derivatives
   - Maximum Entropy Model
   - Naive Bayes
   - CRF
4. Code Review

Parameters

$\boldsymbol{x}^l$	l instance에 대한$\Re^n$ 크기의 feature vector
y^l	l instance에 대한 $\Re$ 크기의 label
w^l	Model의 $\Re^n$크기의 weight vector
b	Model의 $\Re$크기의 bias value

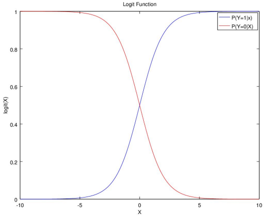
Outline

Logistic Regression Model

Modeling of log of odds ratio

$$\log \frac{P(y^l = 1|x)}{1 - P(y^l = 1|x)} = w^T x + b$$

- $P(y^l = 1|\mathbf{x}^l) = \frac{e^{\mathbf{w}^T \mathbf{x}+b}}{1+e^{\mathbf{w}^T \mathbf{x}+b}}$
- $P(y^l = 0|\mathbf{x}^l) = \frac{1}{1+e^{\mathbf{w}^T \mathbf{x}+b}}$



Outline

Training

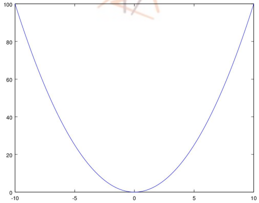
Cost function

$$J = -\sum_{l} y^l \ln\left(P(y^l = 1|\boldsymbol{x}^l)\right) + (1 - y^l)\ln\left(1 - P(y^l = 0|\boldsymbol{x}^l)\right)$$

- If $y^l = 1$
  $\ln\left(P(y^l = 1|\boldsymbol{x}^l)\right)$을 최대화하는 w
- If $y^l = 0$
  $\ln\left(1 - P(y^l = 0|\boldsymbol{x}^l)\right)$을 최대화하는 w

Gradient Descent (because not strictly convex)

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\partial J}{\partial \boldsymbol{w}}$$



Outline

Testing

- $P(y^l = 1|\boldsymbol{x}^l) = \frac{e^{\boldsymbol{w}^T\boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^T\boldsymbol{x}+b}}$
- $P(y^l = 0|\boldsymbol{x}^l) = \frac{1}{1+e^{\boldsymbol{w}^T\boldsymbol{x}+b}}$

Example:

- $x_i = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
- $w = \begin{pmatrix} 3 \\ 7 \end{pmatrix}$
- $b = 1$
- $P(y^l = 1|\boldsymbol{x}^l) = \frac{e^{-3}}{1+e^{-3}} = 0.0474$
- $P(y^l = 0|\boldsymbol{x}^l) = \frac{1}{1+e^{-3}} = 0.953$

Thus, x_i is classified as $y^l = 0$

Outline

Maximum Likelihood

Bayes' formula

$$\begin{aligned} \textit{Posterior} &= \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}} \\ P(x|y) &= \frac{P(y|x) \times P(x)}{P(y)} \end{aligned} \tag{1}$$

Assuming i.i.d

$$\begin{aligned} \textit{likelihood}(\mathbf{w}) &= \prod_l P(Y^l = y_k | \mathbf{x}^l, \mathbf{w}) \\ l(\mathbf{w}) &= -\ln \prod_l P(y^l = y_k | \mathbf{x}^l, \mathbf{w}) \\ &= -\sum_l \ln P(y^l = y_k | \mathbf{x}^l) \end{aligned} \tag{2}$$

Outline

Maximum Entropy

- $$P(y = scissors) = 33\%$$
  $$P(y = rock) = 33\%$$
  $$P(y = paper) = 33\%$$

- $$P(y = scissors) = 98\%$$
  $$P(y = rock) = 1\%$$
  $$P(y = paper) = 1\%$$

Entropy : $H = -\sum_i p_i \log_2 p_i$
Conditional Entropy : $H(y|x) = -\sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x)$

Outline

Maximum Entropy Model

Logistic Regression is equivalent to the basic Maximum Entropy model

$$\begin{aligned}\xi(p,\lambda,\gamma) = &-\left(\sum_{x,y}\tilde{p}(x)p(y|x)\log p(y|x)\right)\\ &+\left(\sum_i \lambda_i \sum_{x,y}\tilde{p}(x)p(y|x)f_i(x,y) - \tilde{p}(x,y)f_i(x,y)\right)\\ &+\gamma\left(1-\sum_y p(y|x)\right)\end{aligned} \tag{3}$$

Maximum Entropy Model

$$\frac{\partial \xi}{\partial p(y|x)} = 0$$

Primal Function

$$p(y|x) = exp\left(\sum_i \lambda_i f_i(x,y)\right) exp\left(-\frac{\gamma}{\tilde{p}(x)} - 1\right)$$

Outline

Naive Bayes

Naive Bayes (Generative) vs Logistic Regression (Discriminative)

- Discriminative:

$$argmax_y p(y|x)$$

- Generative:

$$argmax_y p(x|y)p(y)$$

$$P(x_1, x_2, \ldots, x_n|y) = \prod_i P(x_i|y)$$

$$y \leftarrow argmax_k \left(P(Y = y_k) \prod_i P(x_i|Y = y_k) \right)$$

Naive Bayes

Assuming a Gaussian Naive Bayes with parameters $\mu_{ik}, \sigma_{ik}^2, \pi_k$

$$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$

$$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln\left(\frac{1-\pi}{\pi}\right) + \sum_i \left(x_i \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)\right)}$$

Equivalent model assuming

- Infinite i.i.d data
- Naive bayes assumption

Outline

CRF

Basic equation

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_A \Psi_A(\boldsymbol{x}_A, \boldsymbol{y}_A)$$

Definition of psi

$$\Psi_A(\boldsymbol{x}_A, \boldsymbol{y}_A) = \exp\left(\sum_k \theta_{Ak} f_{Ak}(\boldsymbol{x}_A, \boldsymbol{y}_A) \right)$$

HMM implementation of CRF

$$P(y,\boldsymbol{x})=\exp\left\{\sum_t \left(\sum_{i,j\in S} 1_{y_{t-1}=j} 1_{y_t=i} \ln p(y_t|y_{t-1}) + \sum_{i\in S}\sum_{o\in O} 1_{x_t=o} 1_{y_t=i} \ln p(x_t|y_t)\right)\right\}$$

Outline

Code Review

1. Preprocessing
2. Running Algorithm
3. Evaluation