CS229: Machine Learning

Logistic Regression

Xiangliang Zhang

King Abdullah University of Science and Technology

King Abdullah University of Science and Technology

Logistic Regression

Target variable is not quantitative?

- Logistic Regression

  target variable is **categorical (nominal)**, e.g., married, single, divorced

- Ordinal Logistic Regression

  target variable is **ordinal**, e.g., high, medium, low

It is actually doing classification jobs.

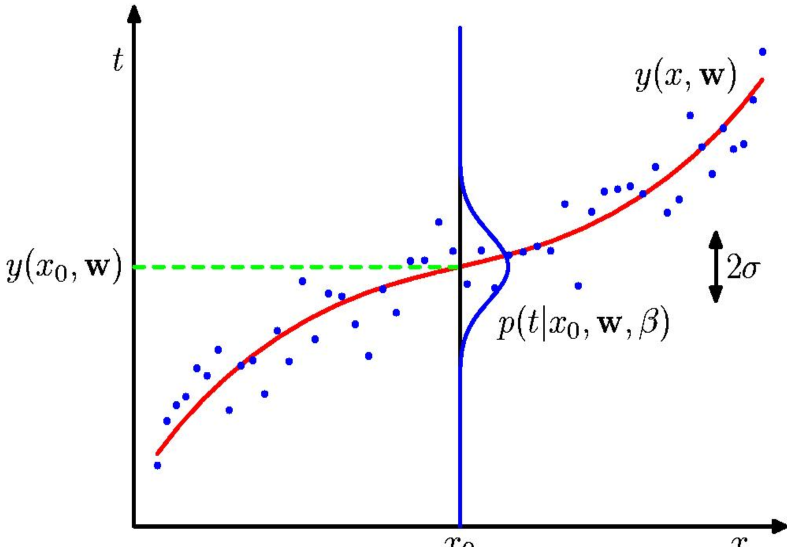
Revisit Regression

Target variable *t* is continuous

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \qquad \text{where} \qquad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

Then,

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$



Logistic Regression – binary variable

Target variable *t* is binary $t \in \{0,1\}$

$$p(t \mid x, \mathrm{w}) = Ber(t \mid \mu(x))$$

where $\mu(x)$ is the parameter of Bernoulli distribution, $p(t = 1 \mid x)$.

Define

$$\mu(x) = \mathrm{sigm}(\mathrm{w}^T x)$$

where $\mathrm{sigm}(a) = \dfrac{1}{1 + e^{-a}}$

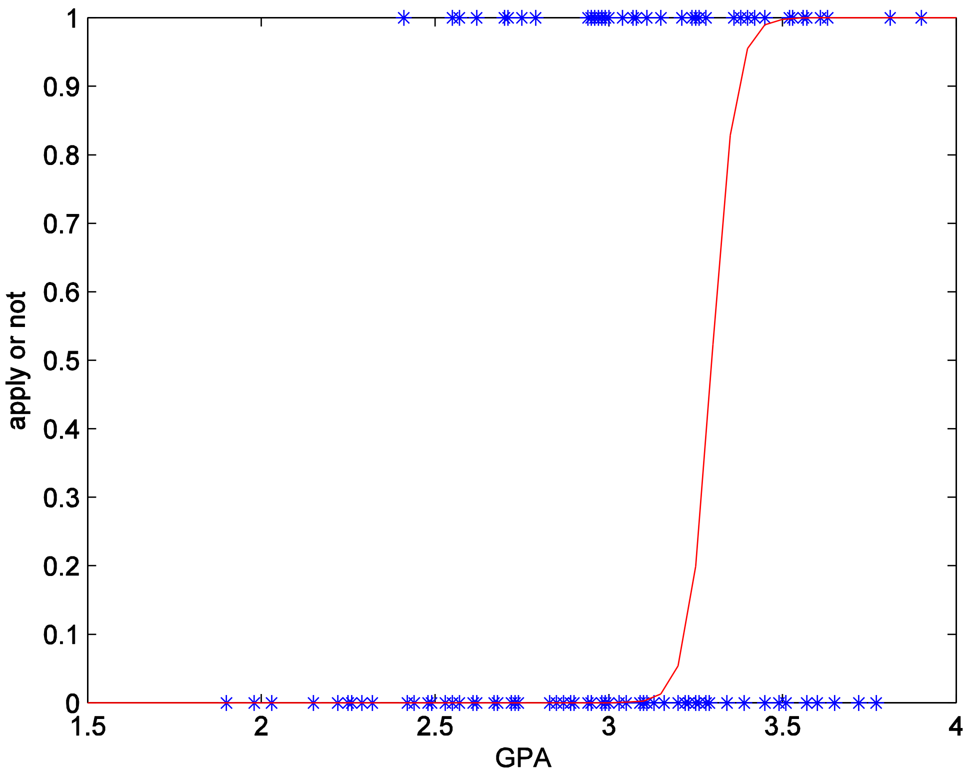
Then,

$$p(t \mid x, \mathrm{w}) = Ber(t \mid \mathrm{sigm}(\mathrm{w}^T x))$$

NOTE: one more dimension with value 1 is added to x.

Logistic Regression – example

apply to graduate school or not



logit and interpretation

The *odds*

$$odds = \frac{p(t=1)}{1 - p(t=1)}$$

Logit: the log of the *odds*

$$\text{logit}(p(t=1)) = \log[\frac{p(t=1)}{1 - p(t=1)}] = ?$$

According to the definition of $\mu(x)$

$$\text{logit}(p(t=1)) = w_0 + w^T x$$

the odds of success is a linear function of x.

Likelihood

Likelihood function

$$L = \prod_{i=1}^{N} \mu_i^{is(t_i=1)} (1-\mu_i)^{is(t_i=0)}$$

$$= \prod_{i=1}^{N} \mu_i^{t_i} (1-\mu_i)^{(1-t_i)}$$

The Negative Log-Likelihood (NLL)

$$NLL(\mathrm{w}) = -\sum_{i=1}^{N} [t_i \ln \mu_i + (1-t_i) \ln(1-\mu_i)]$$

No closed form solution to w.

Gradient of NLL(w)

Derivative of NNL on w,

$$\frac{d\ NLL(\mathrm{w})}{d\ \mathrm{w}} = -\sum_{i=1}^{N}[\frac{t_i}{\mu_i} + \frac{t_i - 1}{1-\mu_i}]\frac{d\mu_i}{d\mathrm{w}}$$

where $$\frac{d\mu_i}{d\mathrm{w}} = \frac{d(1+e^{-\mathrm{w}^T x_i})^{-1}}{d\mathrm{w}} = \mu_i(1-\mu_i)x_i$$

Then, $$\frac{d\ NLL(\mathrm{w})}{d\ \mathrm{w}} = \sum_{i=1}^{N}[\mu_i - t_i]x_i$$

Hessian matrix of NNL(w)

Hessian matrix (second-order partial derivatives)

$$H = \frac{d^2 \ NLL(\mathrm{w})}{d \ \mathrm{w}^2} = \frac{d \ \sum_{i=1}^{N} [\mu_i - t_i] x_i}{d \ \mathrm{w}} = \sum_{i=1}^{N} \frac{d\mu_i}{d\mathrm{w}} x_i^T$$

$$= \sum_{i=1}^{N} \mu_i (1 - \mu_i) x_i x_i^T$$

$$= X^T S X$$

$$S = \begin{bmatrix} \mu_1(1-\mu_1) & \dots & 0 \\ \dots & \ddots & \dots \\ 0 & \dots & \mu_N(1-\mu_N) \end{bmatrix}$$

H is positive definite.

Thus NLL(w) is convex, and has a unique global minimum.

Gradient Descent

Search w* by

$$\mathrm{w}^{k+1} = \mathrm{w}^{k} - \eta g^{k}$$

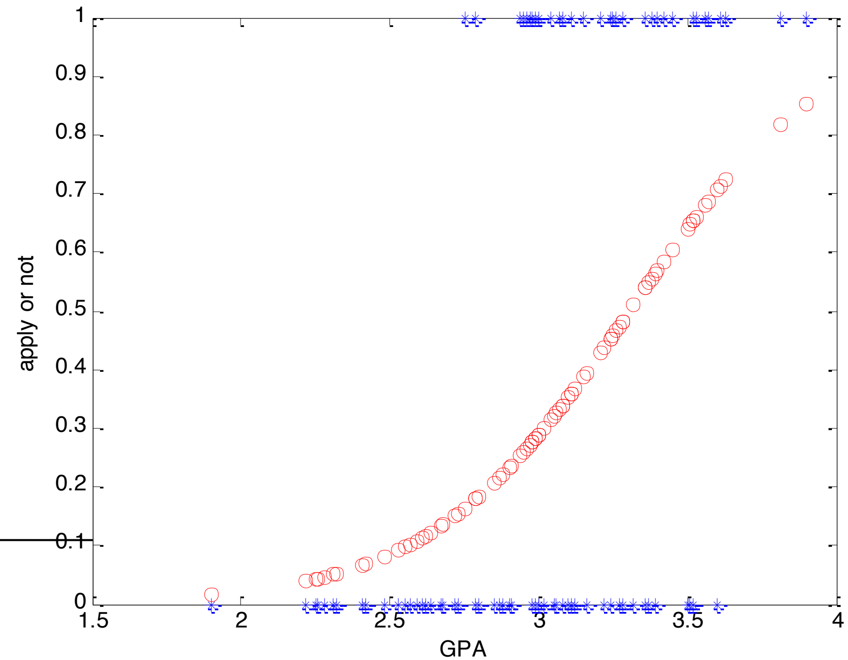
where

$$g^{k} = \frac{d\ NLL(\mathrm{w}^{k})}{d\ \mathrm{w}^{k}} = \sum_{i=1}^{N} [\mu_i - t_i] x_i$$

Prediction

Predict the target

$$t = \begin{cases} 1 & \text{if} \quad \mu = \mathrm{sigm}(\mathrm{w}^T x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$



See a Demo

logit and interpretation

The *odds*

$$odds = \frac{p(t=1)}{1-p(t=1)}$$

Logit: the log of the *odds*

$$\text{logit}(p(t=1)) = \log[\frac{p(t=1)}{1-p(t=1)}] = ?$$

According to the definition of $\mu(x)$

$$\text{logit}(p(t=1)) = w_0 + w^T x$$

the odds of success is a linear function of x.

Multi-class logistic regression

Target variable *t* has *C* nominal values (*C*>2)

$$p(t=c \mid x, \mathrm{W}) = \frac{\exp(\mathrm{w}_{\mathrm{c}}^{\mathrm{T}} x)}{\sum_{c'=1}^{C} \exp(\mathrm{w}_{\mathrm{c'}}^{\mathrm{T}} x)}$$

where columns of W are $\mathrm{w}_{\mathrm{c'}}^{\mathrm{T}}, \; c' = 1 \ldots C$

Multi-class logistic regression - Likelihood

Let $\mu_{ic} = p(t_i = c \mid x_i, \mathrm{W})$ and $t_{ic} = \{0,1\}$ for each *i*, $\sum_{c=1}^{C} t_{ic} = 1$

The likelihood is

$$\prod_{i=1}^{N} \prod_{c=1}^{C} \mu_{ic}^{t_{ic}}$$

The negative log-likelihood (NLL) is

$$NLL(\mathrm{W}) = -\sum_{i=1}^{N} \sum_{c=1}^{C} t_{ic} \log \mu_{ic}$$

$$= -\sum_{i=1}^{N} [\sum_{c=1}^{C} t_{ic} w_c^T x_i - \log \sum_{c'=1}^{C} \exp(w_{c'}^T x_i)]$$

Multi-class logistic regression - Gradient

The gradient of NNL(w) w.r.t. w_c

$$g(w_c) = \frac{\partial NLL(\mathrm{W})}{\partial w_c}$$

$$= \sum_{i=1}^{N} [\frac{\exp(w_c^T x_i)}{\sum_{c'=1}^{C} \exp(w_{c'}^T x_i)} x_i - t_{ic} x_i]$$

$$= \sum_{i=1}^{N} [\mu_{ic} - t_{ic}] x_i$$

Multi-class logistic regression - Hessian

The Hessian matrix has a submatrix (one d*d block)

$$H_{ck} = \frac{dg(w_c)}{\partial w_k} = \sum_{i=1}^{N} \frac{d\mu_{ic}}{dw_k} x_i^T$$

since $\frac{d\mu_{ic}}{dw_k} = \begin{cases} \mu_{ic}(1-\mu_{ic})x_i & \text{when } c = k \\ -\mu_{ic}\mu_{ik}x_i & \text{when } c \neq k \end{cases}$

Then

$$H_{cc} = \sum_{i=1}^{N} \mu_{ic}(1-\mu_{ic})x_i x_i^T$$

$$H_{ck} = \sum_{i=1}^{N} -\mu_{ic}\mu_{ik}x_i x_i^T$$

$$H = \begin{bmatrix} H_{11} & \cdots & H_{1C} \\ \vdots & \ddots & \vdots \\ H_{C1} & \cdots & H_{CC} \end{bmatrix}$$

Positive definite H → NLL(W) has unique minimum

Multi-class logistic regression – learning and prediction

Learn by Gradient descent, for each w_c

$$\mathrm{w}_{\mathrm{c}}^{\mathrm{k+1}} = \mathrm{w}_{\mathrm{c}}^{\mathrm{k}} - \eta g^{k}(\mathrm{w}_{c})$$

where

$$g^{k}(w_c) = \sum_{i=1}^{N} [\mu_{ic} - t_{ic}] x_i$$

Predict the target label for x_i by

$$t_i = \arg\max_{c} \{\mu_{ic}\} = \arg\max_{c} \{\frac{\exp(\mathrm{w}_{\mathrm{c}}^{\mathrm{T}} x_i)}{\sum_{c'=1}^{C} \exp(\mathrm{w}_{\mathrm{c'}}^{\mathrm{T}} x_i)}\}$$