Logistic Regression

Aarti Singh

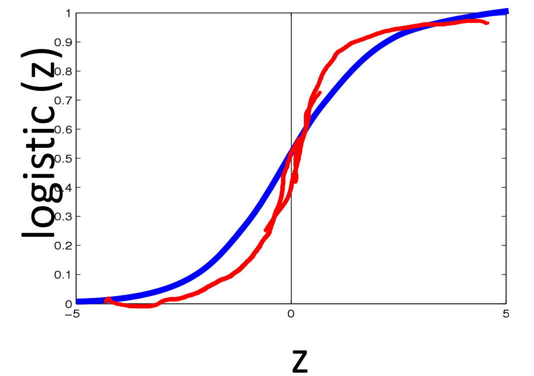
Machine Learning 10-315
Sept 23, 2020

Logistic Regression

Binary classes

Logistic function (or Sigmoid):

$$\frac{1}{1 + exp(-z)}$$



Assumes functional form for P(Y|X):

$$P(Y = 0|X) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

Implies linear decision boundary

bias

How to learn the parameters w_0, w_1, … w_d? (d features)

Maximum (Conditional) Likelihood Estimates

$$\hat{\mathbf{w}}_{MCLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{n} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

Maximum (Conditional) A Posterior Estimates

$w \sim N(0, \kappa^2)$

$$\hat{\mathbf{w}}_{M(C)AP} = \arg\max_{\mathbf{w}} \prod_{j=1}^{n} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})\ P(\mathbf{w})$$

Gradient Ascent for M(C)LE

$$W = \begin{bmatrix} w_0 \\ \vdots \\ w_d \end{bmatrix}$$

Gradient ascent rule for w_0:

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \left.\frac{\partial l(\mathbf{w})}{\partial w_0}\right|_t$$

log likelihood

$$l(\mathbf{w}) = \sum_j \left[y^j (w_0 + \sum_i^d w_i x_i^j) - \ln(1 + exp(w_0 + \sum_i^d w_i x_i^j)) \right]$$

$$\left.\frac{\partial l(w)}{\partial w_0}\right|_{w_1} = \sum_j \left[y^j x_1^j - \underbrace{\frac{\exp(w_0 + \sum_i^d w_i x_i^j)}{1 + \exp(w_0 + \sum_i^d w_i x_i^j)}}_{P(y^j = 1 \mid x_i^j, w)} x_1^j \right]$$

Gradient Ascent for M(C)LE

Gradient ascent rule for w_0:

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \left. \frac{\partial l(\mathbf{w})}{\partial w_0} \right|_t$$

$$l(\mathbf{w}) = \sum_j \left[y^j (w_0 + \sum_i^d w_i x_i^j) - \ln(1 + exp(w_0 + \sum_i^d w_i x_i^j)) \right]$$

$$\frac{\partial l(\mathbf{w})}{\partial w_0} = \sum_j \left[y^j - \frac{1}{1 + exp(w_0 + \sum_i^d w_i x_i^j)} \cdot exp(w_0 + \sum_i^d w_i x_i^j) \right]$$

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \sum_j [y^j - \widehat{P}(Y^j = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]$$

Gradient Ascent for M(C)LE
Logistic Regression

Gradient ascent algorithm: iterate until change < ε

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \sum_j [y^j - \hat{P}(Y^j = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]$$

For i=1,…,d,

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \sum_j x_i^j [y^j - \underbrace{\hat{P}(Y^j = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})}_{\text{Predict what current weight thinks label Y should be}}]$$

repeat

- Gradient ascent is simplest of optimization approaches
  - e.g., Newton method, Conjugate gradient ascent, IRLS (see Bishop 4.3.3)

M(C)AP – Gradient

$\ln p(w) = \sum_i \frac{-w_i^2}{2\kappa^2} + \ln \frac{1}{\kappa\sqrt{2\pi}}$

- Gradient

likelihood

$$\frac{\partial}{\partial w_i} \ln \left[p(\mathbf{w}) \prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]$$

$$p(\mathbf{w}) = \prod_i \frac{1}{\kappa\sqrt{2\pi}} \; e^{\frac{-w_i^2}{2\kappa^2}}$$

Zero-mean Gaussian prior

$\frac{\partial \ln p(w)}{\partial w_1} \propto -\frac{w_1}{\kappa^2}$

$$\frac{\partial}{\partial w_i} \ln p(\mathbf{w}) + \underbrace{\frac{\partial}{\partial w_i} \ln \left[\prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]}_{\text{Same as before}}$$

$$\propto \frac{-w_i}{\kappa^2}$$

Extra term Penalizes large weights

$\lambda (w_i - w_i^0)^2$

Penalization = Regularization

M(C)LE vs. M(C)AP

- Maximum conditional likelihood estimate

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ln \left[\prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]$$

$$\boxed{w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \sum_j x_i^j [y^j - P(Y = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]}$$

- Maximum conditional a posteriori estimate

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ln \left[p(\mathbf{w}) \prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]$$

$$\boxed{w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \left\{ -\frac{1}{\kappa^2} w_i^{(t)} + \sum_j x_i^j [y^j - P(Y = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})] \right\}}$$

Logistic Regression for more than 2 classes

- Logistic regression in more general case, where $Y \in \{y_1,\dots,y_K\}$

$(w_0 \dots w_d)_{K-1}$

for $k<K$

$$P(Y=y_k|X) = \frac{\exp(w_{k0} + \sum_{i=1}^{d} w_{ki}X_i)}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji}X_i)}$$

for $k=K$ (normalization, so no weights for this class)

$$P(Y=y_K|X) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji}X_i)}$$

Predict $f^*(x) = \arg\max_{Y=y} P(Y=y|X=x)$

Is the decision boundary still linear?