Classification with Logisti

Question 1

Why can not we directly use outputs of linear transformation for classification?

1. Scale of original outputs are **not lying in range of 0 to 1**, which does not meet the limitation of possibility
2. Hard-code classificaiton like {0, 1} is **not suitable for derivative,** so is hard for optimization
3. Original outputs may **include negative values**

Solution: Continuous function transformation : Possibility

$$f_\theta(x) = \sigma(\theta^T x)$$

when $K = 2$ get

Logistic (sigmoid)

- Transformation Algorithms: $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$
- Advantage:
  1. symmetry at point (0, 0.5)
  2. sensitive to change of x
  3. continuous, monotonically increasing, derivable

Softmax

- Transformation Algorithms: $\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$
- Advantage:
  1. not negative output
  2. suitable for multi-classify
  3. sum up each equals to 1

Question 2

How can we choose loss functions? And How can we understand them?

Solution: Maximum Likelihood, Cross Entropy

Persepctive of statistic/possibility

Maximum likelihood estimation

$$L(\theta) = \prod_{i=1}^{N} f_\theta(x_i)^{y_i}(1 - f_\theta(x_i))^{1-y_i}$$

Production is hard to optimize and derive

log

Log-likelihood

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{N}[y_i \log f_\theta(x_i) + (1 - y_i)\log(1 - f_\theta(x_i))]$$

Summary is easy to optimize and it is monotonically increasing

~ the same to minus ~

Persepctive of entropy

Cross entropy

$$\begin{aligned} H(p,q) &= -\sum_{i=1}^{n} p(X_i)\log q(X_i) \\ &= -p(X=1)\log q(X=1) - p(X=0)\log q(X=0) \\ &= -y\log f_\theta(x) - (1-y)\log(1 - f_\theta(x)) \end{aligned}$$

To minimize relative entropy is to **minimize H(p,q)**

relative entropy

$$D_{KL}(p \,||\, q) = E_{X\sim p(x)}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] = \sum_{i=1}^{n}(p(X_i)\log p(X_i) - p(X_i)\log q(X_i)) = -H(p) + H(p,q)$$

KL divergence = average distance between p and q distribution with p possibility as weights

Distance between p and q distribution

Information/Entropy

information → $I(X_i) = -\log P(X_i)$ → Entropy → $H(p) = E_{X\sim p(x)}[I(X)] = \sum_{i=1}^{n} P(X_i)I(X_i) = -\sum_{i=1}^{n} P(X_i)\,l$

ic and Softmax transformation

Question 3

How can we evaluate classified results?

Solution:
Confusion Matrix

Change with threshold

- Precision
  - Precise enough?
  - $Prec = \frac{TP}{TP+FP}$
  - Prediceted and actually true positive / all predicted true positive
- Recall
  - Inculde enough?
  - $Rec = \frac{TP}{TP+FN}$
  - Prediceted and actually true positive / all actually true positive
- F1
  - Harmonic mean:
  - $F_1 = \frac{2}{\frac{1}{Prec} + \frac{1}{Rec}}$
  - Weighted mean:
  - $F_\beta = (1+\beta^2) \times \frac{2 \times Prec \times Rec}{\beta^2 \times Prec + Rec}$

Not change with threshold

- TPR
  - Equals to Recall
  - $TPR = \frac{TP}{TP+FN}$
- FPR
  - False positive rate
  - $FPR = \frac{FP}{FP+TN}$
- ROC/AUC

1. more clearly model diferentiates classes , the more bigger space of AUC
2. AUC values is not influenced by threshold

GLM, Generalized Linear Model

Exponential distribution family

$$P(y \mid \eta) = b(y)e^{\eta^T T(y) - A(\eta)}$$

Distribution

Gaussian Distribution

- $b(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}$
- $T(y) = y$
- $A(\eta) = \frac{\eta^2}{2}$

Bernoulli Distribution

- $b(y) = 1$
- $T(y) = y$
- $A(\eta) = \ln(1+e^\eta)$

Expectation / Variance

$E(y \mid \eta) = \nabla_\eta A(\eta)$

$Vary \mid \eta) = \nabla_\eta^2 A(\eta)$

$f_\theta(X) = \nabla_\eta A(\eta)$

$= \nabla_\eta \left(\frac{\eta^T \eta}{2} \right) = \eta = X\theta$

the original linear model

$E(y \mid \eta) = \nabla_\eta A(\eta)$

$f_\theta(X) = \nabla_\eta A(\eta)$

$= \nabla_\eta \ln(1+e^\eta)$

$= \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-X\theta}}$

$= \sigma(X\theta)$

sigmoid transformed linear model

$\log P(X_i)$