Advanced Machine Learning

John Min

March 24, 2014

1 Logistic Regression

Recall the Bernoulli random variable: for $x \in \{0,1\}, \mu$ probability of heads:

$$p(x|\mu) = \mu^x (1-\mu)^{1-x}$$

$$p(x|\mu) = (1-\mu) \exp\left\{ x \ln\left(\frac{\mu}{1-\mu}\right) \right\}$$

$\gamma = \text{logit}(\mu)$ The inverse transform for the logit is the sigmoid:

$$\sigma(\gamma) = \frac{1}{1+\exp(-\gamma)}$$

1.1 Logistic Regression vs. Least Squares

- One can formulate LS classification, modeling each C_k with its own linear mode, and minimizing the squared error between predicted and observed labels
- However, LS is not robust with respect to outliers
- Heavy tailed modeling?

2 Neural Nets: Data-Adaptive Learning

The limitation of the GLM modeling framework comes from its simplicity that facilitates model fitting. The response is modeled as $y = \gamma(w^\top x)$ with γ being some typically monotonic transformation.

- logit/sigmoid (Bernoulli, Multinomial)
- log/exp (Poisson)
- $1/x$ (Gamma)

Can we learn a more complex predictive mechanism?

$$y = f(x)$$

- Parametric form: formulate class of functions (e.g. polynomials, cubic splines) and learn their coefficients.
- Non-parametric: recover functions from inputs to outputs, penalizing complexity in functional representation.
- Data-adapted: formulate a mechanism, and learn the 'knobs' that configure it to input/output information (NN)

2.1 Activation Functions

2.1.1 Sigmoids

Sigmoids $\sigma(x) = \frac{1}{1+exp(-\gamma x)}$ are widely used as activation functions:

- Small γ give linear-like activation, reducing the NN to a convex model
- Large γ gives a step-function, corresponding to the perceptron

2.2 Training the NN

Given the input/output pair $(x, \bar{y})$, the predicted output is $y = f(z)$, a function of hidden units. Training is performed using cross-entropy.

$$f(z) = \begin{bmatrix} \sigma\left(v_1^\top z - \xi_1\right) \\ \vdots \\ \sigma\left(v_k^\top z - \xi_k\right) \end{bmatrix}$$

We need to learn V, Ξ by using the soft-max:

$$\min_{V,\xi} \ln\left(\sum_{j=1}^{k} \exp\left(v_j^\top z - \xi_j\right)\right) - \left(v_p^\top z - \xi_p\right)$$