Intro to LR Opt

qiyiping@meituan.com

Outline

- Optimization Basics
- Logistic regression
- Reference

Optimization Basics

$$\min f(x)$$
$$subject\ to\ x \in X$$

- Unconstrained Optimization ($X = R^n$)
  - Local/Global minima
$$f(x^*) \le f(x), \forall x\ with\ |x - x^*| \le \epsilon$$
$$f(x^*) \le f(x), \forall x \in R^n$$
  - Necessary condition for differentiable f(x)
$$\nabla f(x^*) = 0$$
  - No distinction between local/global minima for convex cost function
- Optimization Over a Convex Set ($X \subset R^n, X\ is\ convex$)
  - Necessary condition for local optimal
$$\nabla f(x^*)(x - x^*) \ge 0, \forall x \in X$$

Logistic Regression

- Generalized Linear model

$$E(y) = g^{-1}(w^T x)$$

- Logit link function

$$g^{-1}(z) = \sigma(z) = \frac{1}{1 + \exp(-z)}$$

- Logistic regression model

$$P(y = \pm 1 | x, w) = \sigma(y w^T x) = \frac{1}{1 + \exp(-y w^T x)}$$

- MAP (Unconstrained Convex optimization problem)

$$\min_{w \in R^n} f(X, Y, w) = \sum_{i=1}^{N} \log(1 + \exp(-y_i w^T x_i)) + \lambda R(w)$$

Logistic Regression cont'd

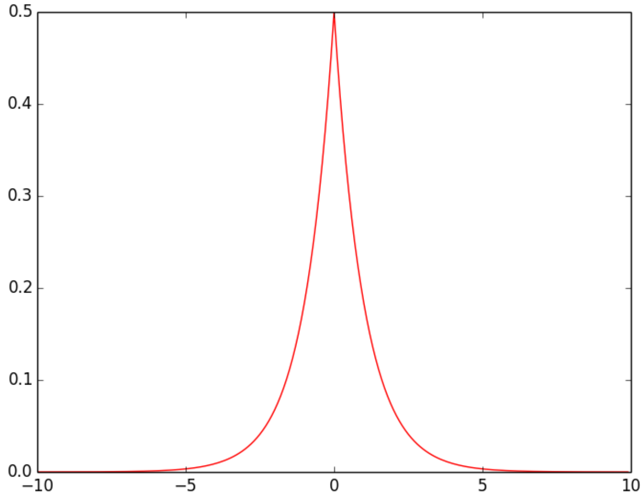
- Why is LR widely used? ⇔ Why sigmoid?
  - For the two-class classification problem, the posterior probability of class C can be written as a logistic sigmoid acting on a linear function of x, for a wide choice of class-conditional distributions p(x|C)
  - E.g. p(x|C) is Gaussian distributed

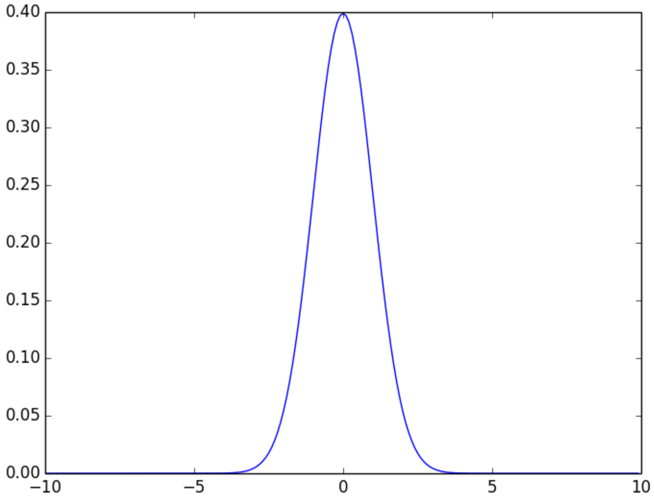
L1/L2 regularization

- Prior distribution of parameter w
  - Gaussian prior => L2 regularization
  - Laplace prior => L1 regularization

L1/L2 regularization cont'd

- L1 yields sparse models
  - Laplace prior likes to shrink coefficients to zero
- L1 based feature selection
  - Bayesian feature selection





Gradient Descent

- First order Taylor series expansion
  $$f(x + \Delta x) = f(x) + \nabla f(x)^T \Delta x + o(|\Delta x|)$$
- Descent direction d
  $$\nabla f(x)^T d < 0$$
- Many descent direction specified in the form
  $$d = -D\nabla f(x)$$

where D is positive definite

Gradient Descent cont'd

- Steepest Descent

$$D = I$$
$$d = -\nabla f(x)$$
$$\Delta x = -\alpha \Delta f(x)$$

- Good/Bad Example
- Step size selection
  - Minimization rule
  - Limited minimization rule
  - Successive step size reduction
- Stochastic gradient descent
  - Approximates the true gradient by considering a single training example at a time.
- For LR:

$$d = \sum_i \big(1 - \sigma(y_i w^T x)\big) y_i x_i - \lambda w$$

Newton's Method

- Second order Taylor series expansion

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x$$

$$\nabla f(x) + \nabla^2 f(x) \Delta x = 0$$

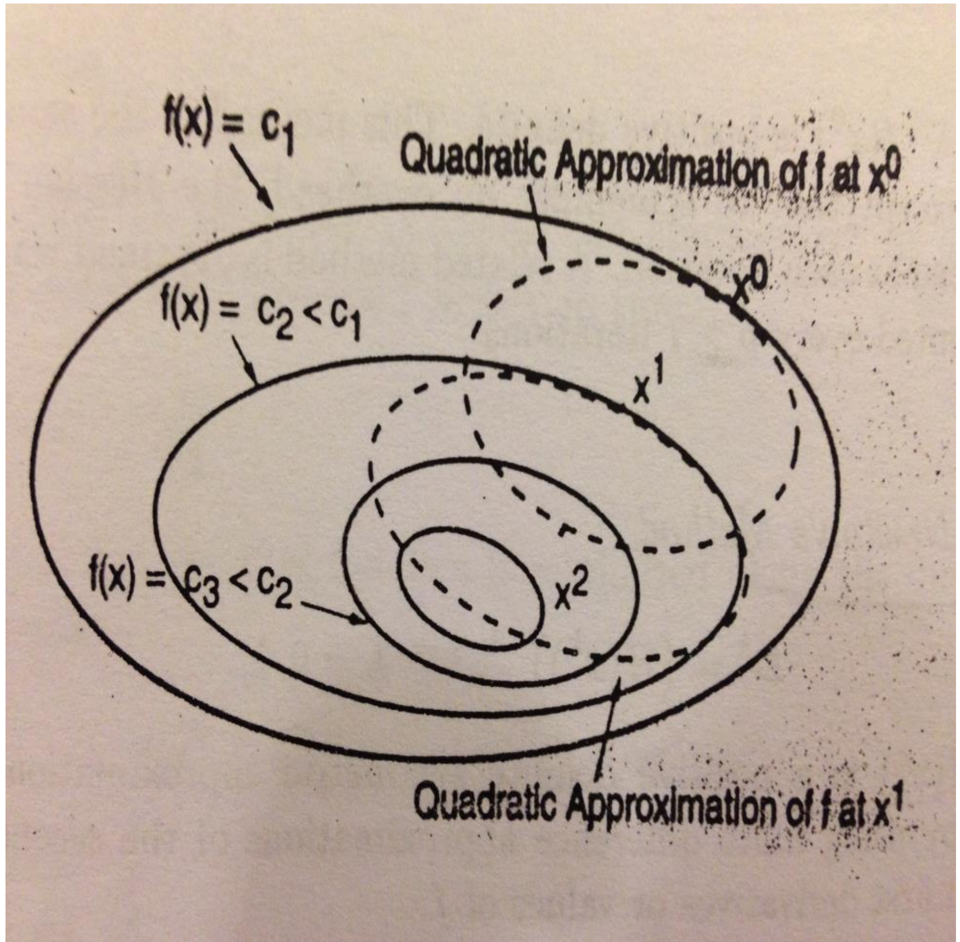
$$\Delta x = -\alpha (\nabla^2 f(x))^{-1} \nabla f(x)$$

- If f(x) is convex, $\nabla^2 f(x)$ is semi-positive definite
- $-(\nabla^2 f(x))^{-1} \nabla f(x)$ is a descent direction
- Converges fast & does not exhibit zig-zagging behavior
- For LR:

$$H = XAX^T + \lambda I$$

where A is a positive definite diagonal matrix.

Newton’s Method cont’d



quasi-Newton Method

- Emulate Newton's method
- Avoid the second derivative calculations
  $$\nabla f(x_{k+1}) - \nabla f(x_k) \approx D_k(x_{k+1} - x_k)$$
- Different methods to calculate D_k and its inverse
  - BFGS
  - SR1

Even more…

- Coordinate Optimization
- FTRL (kdd13)
- …

Reference

- Nonlinear Programming, Dimitri Bertsekas
- Pattern Recognition and Machine Learning, Christopher M. Bishop
- Large-scale Bayesian logistic regression for text categorization, Genkins et al.
- Ad Click Prediction: a View from the Trenches, H. Brendan McMahan, Gary Holt, D. Sculley et al.
- http://www.csie.ntu.edu.tw/~cjlin/liblinear/