10-701 Recitation:
Loss, Regularization, and Dual*

Jay-Yoon Lee

02/26/2015

*Adopted figures from 10725 lecture slides and from the book 'Elements of Statistical Learning'

Loss and Regularization

- Optimization problem can be expressed as to minimize "Loss".

$$\underset{\text{models } M}{\arg\min} \sum_{i=1}^{n} \ell(x_i; M)$$

- If want to maximize your "objective function", negative of objective function is loss.

Loss and Regularization

- Optimization problem can be expressed as to minimize "Loss".

$$\underset{\text{models } M}{\arg\min} \sum_{i=1}^{n} \ell(x_i; M)$$

- Introduce "Regularization" term (or "penalty") to prevent overfitting or satisfy constraints

$$\Longrightarrow \underset{\text{models } M}{\arg\min} \sum_{i=1}^{n} \ell(x_i; M) + \text{penalty}(M)$$

Loss and Regularization

- Example: “Loss” of linear regression problem

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2$$

Loss and Regularization

- Example: "Loss" of linear regression problem

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2$$

- Example: "Penalty" of linear regression

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + ||\beta||_1$$

Loss and Regularization

- More Examples

Model	Fit measure	Entropy measure[4][5]
AIC/BIC	$\|Y - X\beta\|_2$	$\|\beta\|_0$
Ridge regression	$\|Y - X\beta\|_2$	$\|\beta\|_2$
Lasso[6]	$\|Y - X\beta\|_2$	$\|\beta\|_1$
Basis pursuit denoising	$\|Y - X\beta\|_2$	$\lambda\|\beta\|_1$
Rudin-Osher-Fatemi model (TV)	$\|Y - X\beta\|_2$	$\lambda\|\nabla\beta\|_1$
Potts model	$\|Y - X\beta\|_2$	$\lambda\|\nabla\beta\|_0$
RLAD[7]	$\|Y - X\beta\|_1$	$\|\beta\|_1$
Dantzig Selector[8]	$\|X^\top(Y - X\beta)\|_\infty$	$\|\beta\|_1$
SLOPE[9]	$\|Y - X\beta\|_2$	$\sum_{i=1}^{p} \lambda_i \|\beta\|_{(i)}$

From wikipedia: http://en.wikipedia.org/wiki/Regularization_(mathematics)

Dual: Lagrangian Function

- Many constrained optimization can be expressed in term of "loss" and "penalty".

- Recall Lagrangian function
  - Primal $$\underset{x}{\text{minimize}}\, f(x) \text{ subject to } c_i(x) \leq 0$$
  - Dual $$\underset{\alpha}{\text{maximize}}\, L(x(\alpha), \alpha)$$

$$L(x, \alpha) = f(x) + \sum_i \alpha_i c_i(x)$$

Dual: Lagrangian Function

- More generally,

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, \quad i = 1, \dots m \\ & \ell_j(x) = 0, \quad j = 1, \dots r \end{aligned}$$

- Lagrangian

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$$

From 10725 Lecture notes

Dual: Lagrangian Function

- Important Property
  - Lagrangian function is lower bound of loss function.

Important property: for any $u \geq 0$ and v,

$$f(x) \geq L(x,u,v) \quad \text{at each feasible } x$$

Why? For feasible x,

$$L(x,u,v) = f(x) + \sum_{i=1}^{m} u_i \underbrace{h_i(x)}_{\leq 0} + \sum_{j=1}^{r} v_j \underbrace{\ell_j(x)}_{=0} \leq f(x)$$

From 10725 Lecture notes

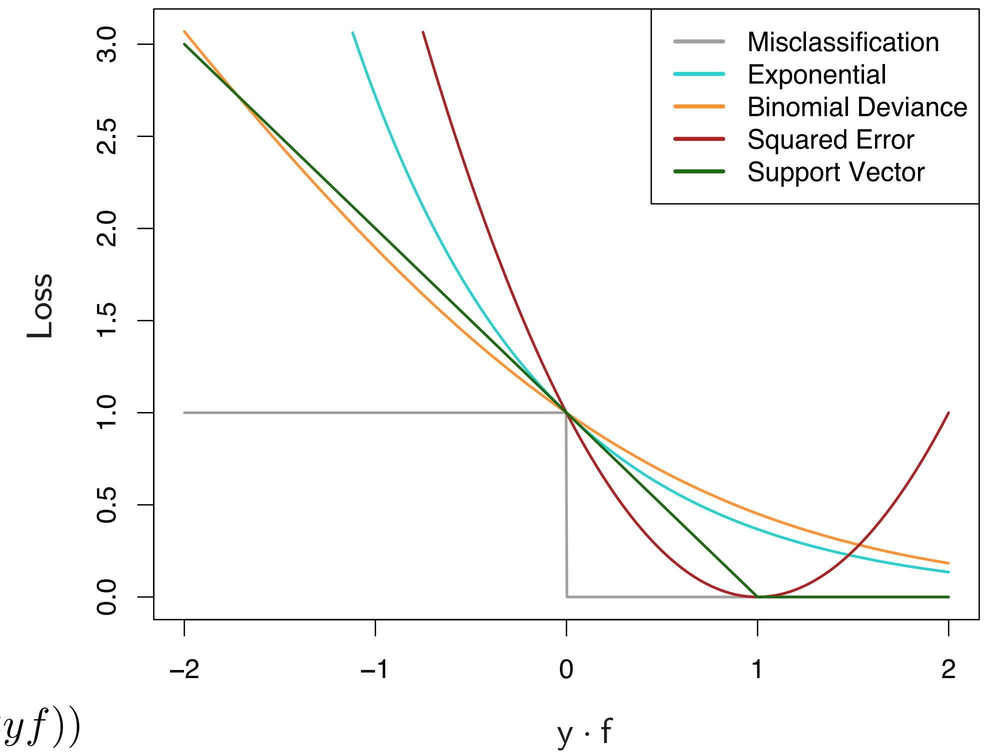
Loss Functions (Classification)

- Model

$$\begin{array}{rl} \text{Model :} & f \\ \text{Label :} & y = \pm 1 \\ \text{Prediction:} & \text{sign}(f) \end{array}$$

- Loss function

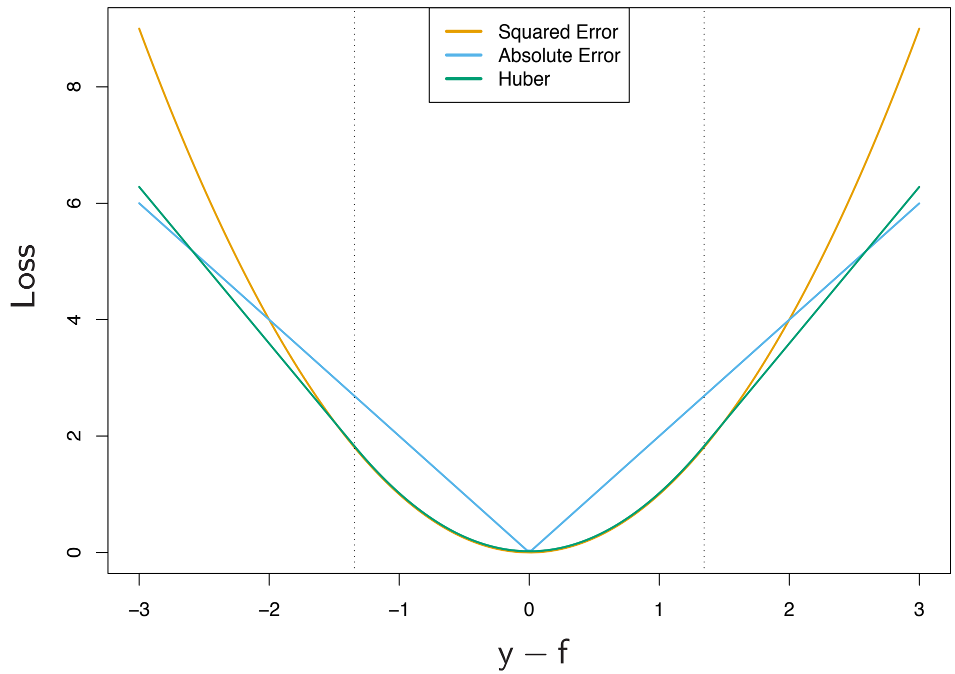
$$\begin{array}{cl} \text{misclassification (0-1)} & I(\text{sign}(f \neq y)) \\ \text{exponential} & \exp(-yf) \\ \text{binomail deviance} & \log(1+\exp(-2yf)) \\ \text{hinge} & max(1-yf, 0) \end{array}$$



From Elements of Statistical Learning, 2nd edition, Springers

Loss Functions (Regression)

- Loss



$$\begin{array}{ll} \text{Squared-Error} & \ell(y, f(x)) = (y - f(x))^2 \\ \text{Absolute Loss} & \ell(y, f(x)) = |y - f(x)| \\ \text{Huber Loss} & \ell(y, f(x)) = \begin{cases} (y - f(x))^2 & \text{for } |y - f(x)| \leq \delta \\ 2\delta|y - f(x)| - \delta^2 & \text{otherwise.} \end{cases} \end{array}$$

From Elements of Statistical Learning, 2nd edition, Springers

Loss Functions

- Classification

misclassification (0-1)	$I(\text{sign}(f \neq y))$
exponential	$\exp(-yf)$
binomail deviance	$\log(1+\exp(-2yf))$
hinge	$max(1-yf, 0)$

- Regression

Squared-Error	$\ell(y, f(x)) = (y - f(x))^2$
Absolute Loss	$\ell(y, f(x)) = \|y - f(x)\|$
Huber Loss	$\ell(y, f(x)) = \begin{cases} (y-f(x))^2 & \text{for } \|y-f(x)\| \leq \delta \\ 2\delta\|y-f(x)\| - \delta^2 & \text{otherwise.} \end{cases}$

From Elements of Statistical Learning, 2nd edition, Springers

Classification Examples

Linear soft margin problem

$$\underset{w,b}{\text{minimize}} \ \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i\left[\langle w, x_i\rangle + b\right] \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

Dual problem

$$\underset{\alpha}{\text{maximize}} - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_i \alpha_i$$

$$\text{subject to } \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \in [0, C]$$

From 701 lecture notes

Classification Examples

- Logistic Regression

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

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

$$\begin{aligned} l(W) &\equiv \ln \prod_l P(Y^l|X^l,W) \\ &= \sum_l Y^l(w_0 + \sum_i^n w_i X_i^l) - \ln(1 + exp(w_0 + \sum_i^n w_i X_i^l)) \end{aligned}$$

Penalty Functions

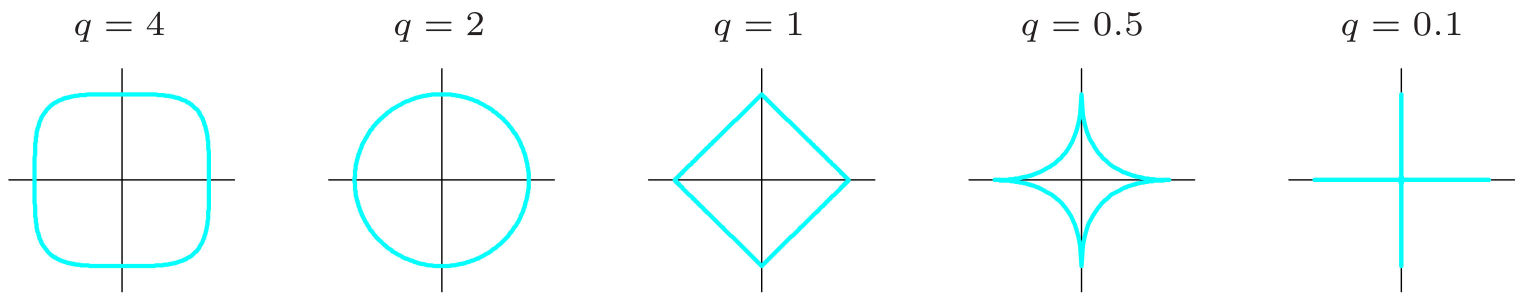
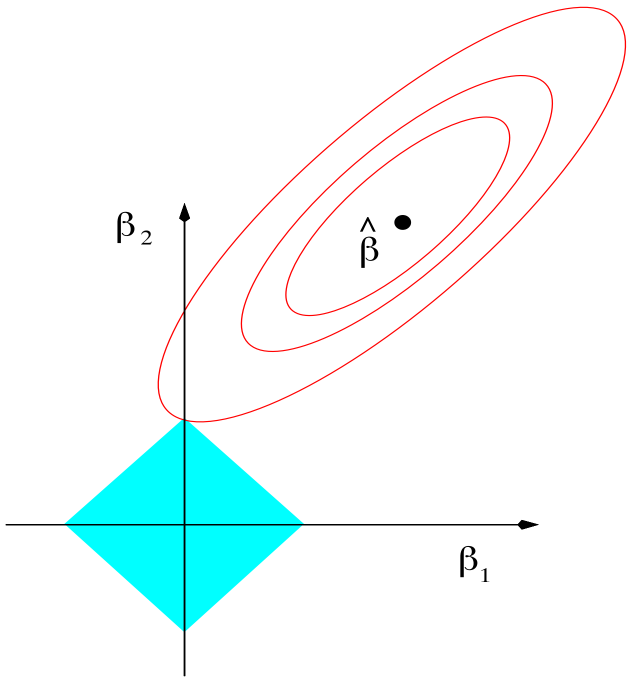
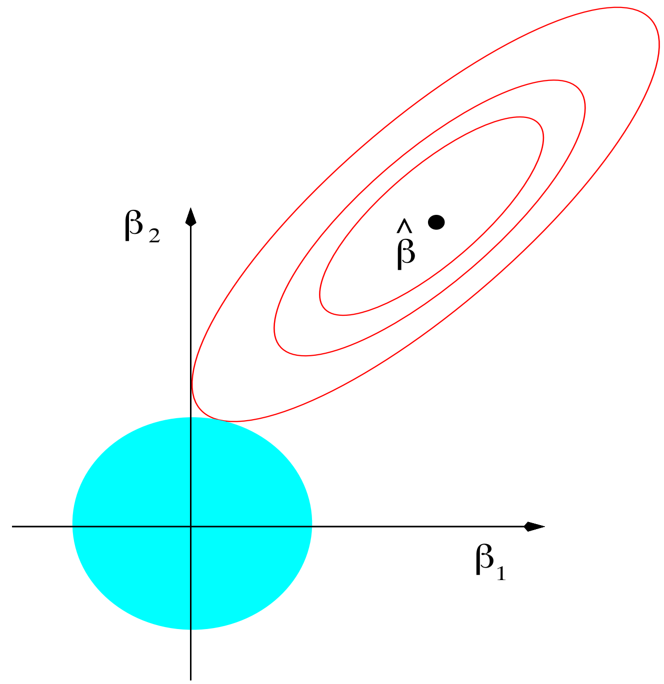


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

From Elements of Statistical Learning, 2^{nd} edition, Springers

Penalty Functions





LASSO

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + ||\beta||_1$$

Ridge Regression

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + ||\beta||_2$$

From Elements of Statistical Learning, 2nd edition, Springers

Back up slides

Lagrange Multipliers

From 10701 Lecture 5

- Lagrange function

$$L(x, \alpha) := f(x) + \sum_{i=1}^{n} \alpha_i c_i(x) \text{ where } \alpha_i \geq 0$$

- Saddlepoint Condition
  If there are x* and nonnegative α* such that

$$L(x^*, \alpha) \leq L(x^*, \alpha^*) \leq L(x, \alpha^*)$$

then x* is an optimal solution to the constrained optimization problem

Necessary Kuhn-Tucker Conditions

From 10701 Lecture 5

- Assume optimization problem
  - satisfies the constraint qualifications
  - has convex differentiable objective + constraints
- Then the KKT conditions are necessary & sufficient

$$\partial_x L(x^*, \alpha^*) = \partial_x f(x^*) + \sum_i \alpha_i^* \partial_x c_i(x^*) = 0 \text{ (Saddlepoint in } x^*)$$

$$\partial_{\alpha_i} L(x^*, \alpha^*) = c_i(x^*) \leq 0 \text{ (Saddlepoint in } \alpha^*)$$

$$\sum_i \alpha_i^* c_i(x^*) = 0 \text{ (Vanishing KKT-gap)}$$

Yields algorithm for solving optimization problems
Solve for saddlepoint and KKT conditions

Lagrangian

Consider general minimization problem

$$\min_{x \in \mathbb{R}^n} \; f(x)$$
$$\text{subject to } \; h_i(x) \leq 0, \;\; i = 1, \ldots m$$
$$\ell_j(x) = 0, \;\; j = 1, \ldots r$$

Need not be convex, but of course we will pay special attention to convex case

We define the Lagrangian as

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$$

New variables $u \in \mathbb{R}^m, v \in \mathbb{R}^r$, with $u \geq 0$ (implicitly, we define $L(x, u, v) = -\infty$ for $u < 0$)

From 725lecture notes