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”.

\[
\arg \min_{M} \sum_{i=1}^{n} \ell(x_i; M)
\]

• 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”.  

\[
\arg\min_{\text{models } M} \sum_{i=1}^{n} \ell(x_i; M)
\]

• Introduce “Regularization” term (or “penalty”) to prevent overfitting or satisfy constraints

\[
\implies \arg\min_{\text{models } M} \sum_{i=1}^{n} \ell(x_i; M) + \text{penalty}(M)
\]
Loss and Regularization

• Example: “Loss” of linear regression problem

\[
\arg\min_{\beta} \| y - X\beta \|_2^2
\]
Loss and Regularization

• Example: “Loss” of linear regression problem

\[
\arg \min_{\beta} ||y - X\beta||_2^2
\]

• Example: “Penalty” of linear regression

\[
\arg \min_{\beta} ||y - X\beta||_2^2 + ||\beta||_1
\]
## Loss and Regularization

- **More Examples**

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

[^4][^5]: 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  \[
  \min_{x} f(x) \text{ subject to } c_i(x) \leq 0
  \]
  
  – Dual  \[
  \max_{\alpha} L(x(\alpha), \alpha)
  \]

  \[
  L(x, \alpha) = f(x) + \sum \alpha_i c_i(x)
  \]
Dual: Lagrangian Function

• More generally,

\[
\begin{align*}
\min_{x \in \mathbb{R}^n} & \quad f(x) \\
\text{subject to} & \quad h_i(x) \leq 0, \quad i = 1, \ldots, m \\
& \quad \ell_j(x) = 0, \quad j = 1, \ldots, r
\end{align*}
\]

• 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 \( x \),

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

From 10725 Lecture notes
Loss Functions (Classification)

• Model

  Model : \( f \)
  Label : \( y = \pm 1 \)
  Prediction: \( \text{sign}(f) \)

• Loss function

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

From Elements of Statistical Learning, 2nd edition, Springers
Loss Functions (Regression)

- **Loss**

  - **Squared Error**
    \[ \ell(y, f(x)) = (y - f(x))^2 \]

  - **Absolute Error**
    \[ \ell(y, f(x)) = |y - f(x)| \]

  - **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
Loss Functions

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

- **Regression**
  
<table>
<thead>
<tr>
<th>Loss Type</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squared-Error</td>
<td>( \ell(y, f(x)) = (y - f(x))^2 )</td>
</tr>
<tr>
<td>Absolute Loss</td>
<td>( \ell(y, f(x)) =</td>
</tr>
<tr>
<td>Huber Loss</td>
<td>( \ell(y, f(x)) = \begin{cases} (y - f(x))^2 &amp; \text{for }</td>
</tr>
</tbody>
</table>

From Elements of Statistical Learning, 2nd edition, Springers
Classification Examples

Linear soft margin problem

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \\
\text{subject to} & \quad y_i [\langle w, x_i \rangle + b] \geq 1 - \xi_i \text{ and } \xi_i \geq 0
\end{align*}
\]

Dual problem

\[
\begin{align*}
\text{maximize} & \quad -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i \\
\text{subject to} & \quad \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \in [0, C]
\end{align*}
\]

From 701 lecture notes
Classification Examples

- Logistic Regression

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

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

\[
l(W) \equiv \ln \prod_l P(Y^l|X^l, W)
\]

\[
= \sum_l Y^l(w_0 + \sum_i^n w_i X^l_i) - \ln(1 + \exp(w_0 + \sum_i^n w_i X^l_i))
\]
Penalty Functions

\[
\beta_1^2 + \beta_2^2 \leq t, \quad \text{while that for lasso is } |\beta_1| + |\beta_2| \leq t.
\]

Both methods find the first point where the elliptical contours hit the constraint region. Unlike the disk, the diamond has corners; if the solution occurs at a corner, then it has one parameter \( \beta_j \) equal to zero. When \( p > 2 \), the diamond becomes a rhomboid, and has many corners, flat edges and faces; there are many more opportunities for the estimated parameters to be zero.

We can generalize ridge regression and the lasso, and view them as Bayes estimates. Consider the criterion

\[
\tilde{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}
\]

for \( q \geq 0 \). The contours of constant value of \( \sum_j |\beta_j|^q \) are shown in Figure 3.12, for the case of two inputs. Thinking of \( |\beta_j|^q \) as the log-prior density for \( \beta_j \), these are also the equi-contours of the prior distribution of the parameters. The value \( q = 0 \) corresponds to variable subset selection, as the penalty simply counts the number of nonzero parameters; \( q = 1 \) corresponds to the lasso, while \( q = 2 \) to ridge regression. Notice that for \( q \leq 1 \), the prior is not uniform in direction, but concentrates more mass in the coordinate directions. The prior corresponding to the \( q = 1 \) case is an independent double exponential (or Laplace) distribution for each input, with density \( \frac{1}{2 \tau} \exp(-|\beta|/\tau) \) and \( \tau = 1/\lambda \).

The case \( q = 1 \) (lasso) is the smallest \( q \) such that the constraint region is convex; non-convex constraint regions make the optimization problem more difficult.

In this view, the lasso, ridge regression and best subset selection are Bayes estimates with different priors. Note, however, that they are derived as posterior modes, that is, maximizers of the posterior. It is more common to use the mean of the posterior as the Bayes estimate. Ridge regression is also the posterior mean, but the lasso and best subset selection are not.

Looking again at the criterion (3.53), we might try using other values of \( q \) besides 0, 1, or 2. Although one might consider estimating \( q \) from the data, our experience is that it is not worth the effort for the extra variance incurred. Values of \( q \in (1, 2) \) suggest a compromise between the lasso and ridge regression. Although this is the case, with \( q > 1 \), \( |\beta_j|^q \) is differentiable at 0, and so does not share the ability of lasso (\( q = 1 \)) for \( q = 4 \) \( q = 2 \) \( q = 1 \) \( q = 0.5 \) \( q = 0.1 \).  

\[\text{FIGURE 3.12. Contours of constant value of } \sum_j |\beta_j|^q \text{ for given values of } q.\]
Penalty Functions

\[
\text{LASSO} \quad \arg\min_\beta \| y - X\beta \|^2_2 + \| \beta \|_1
\]

\[
\text{Ridge Regression} \quad \arg\min_\beta \| y - X\beta \|^2_2 + \| \beta \|^2_2
\]

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 \( \alpha^* \) 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

\[
\begin{align*}
\partial_x L(x^*, \alpha^*) &= \partial_x f(x^*) + \sum_i \alpha_i^* \partial_x c_i(x^*) = 0 \quad \text{(Saddlepoint in } x^*) \\
\partial_{\alpha_i} L(x^*, \alpha^*) &= c_i(x^*) \leq 0 \quad \text{(Saddlepoint in } \alpha^*) \\
\sum_i \alpha_i^* c_i(x^*) &= 0 \quad \text{(Vanishing KKT-gap)}
\end{align*}
\]

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

Consider general minimization problem

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

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