Fundamentals of Linear Algebra and Optimization
Ridge Regression

Jean Gallier and Jocelyn Quaintance

CIS Department
University of Pennsylvania
jean@cis.upenn.edu

May 7, 2020
Ridge Regression

The problem of solving an overdetermined or underdetermined linear system \(Aw = y\), where \(A\) is an \(m \times n\) matrix, arises as a “learning problem” in which we observe a sequence of data \(((a_1, y_1), \ldots, (a_m, y_m))\), viewed as input-output pairs of some unknown function \(f\) that we are trying to infer, where the \(a_i\) are the rows of the matrix \(A\) and \(y_i \in \mathbb{R}\).
Ridge Regression

The problem of solving an overdetermined or underdetermined linear system $A w = y$, where $A$ is an $m \times n$ matrix, arises as a “learning problem” in which we observe a sequence of data $((a_1, y_1), \ldots, (a_m, y_m))$, viewed as input-output pairs of some unknown function $f$ that we are trying to infer, where the $a_i$ are the rows of the matrix $A$ and $y_i \in \mathbb{R}$.

The values $y_i$ are sometimes called labels or responses.
Ridge Regression

The problem of solving an overdetermined or underdetermined linear system \( Aw = y \), where \( A \) is an \( m \times n \) matrix, arises as a “learning problem” in which we observe a sequence of data \( ((a_1, y_1), \ldots, (a_m, y_m)) \), viewed as input-output pairs of some unknown function \( f \) that we are trying to infer, where the \( a_i \) are the rows of the matrix \( A \) and \( y_i \in \mathbb{R} \).

The values \( y_i \) are sometimes called *labels* or *responses*.

The simplest kind of function is a linear function \( f(x) = x^\top w \), where \( w \in \mathbb{R}^n \) is a vector of coefficients usually called a *weight vector*, or sometimes an *estimator*. 
Ridge Regression: Least-Squares Solution

Since the problem is overdetermined and since our observations may be subject to errors, we can’t solve for $w$ exactly as the solution of the system $Aw = y$, so instead we solve the least-square problem of minimizing $\|Aw - y\|^2$. In an earlier module we showed that this problem can be solved using the pseudo-inverse.
Ridge Regression: Least-Squares Solution

Since the problem is overdetermined and since our observations may be subject to errors, we can’t solve for $w$ exactly as the solution of the system $Aw = y$, so instead we solve the least-square problem of minimizing $\|Aw - y\|_2^2$.

In an earlier module we showed that this problem can be solved using the pseudo-inverse.
Ridge Regression: Least-Squares Solution

Since the problem is overdetermined and since our observations may be subject to errors, we can’t solve for $w$ exactly as the solution of the system $Aw = y$, so instead we solve the least-square problem of minimizing $\|Aw - y\|_2^2$.

In an earlier module we showed that this problem can be solved using the pseudo-inverse.

We know that the minimizers $w$ are solutions of the normal equations $A^T Aw = A^T y$, but when $A^T A$ is not invertible, such a solution is not unique so some criterion has to be used to choose among these solutions.
Ridge Regression: Least-Squares Solutions

One solution is to pick the unique vector $w^+$ of smallest Euclidean norm $\|w^+\|_2$ that minimizes $\|Aw - y\|_2^2$. The solution $w^+$ is given by $w^+ = A^+ y$, where $A^+$ is the pseudo-inverse of $A$. The matrix $A^+$ is obtained from an SVD of $A$, say $A = V U^\top$. Namely, $A^+ = U^+ V^\top$, where $^+$ is the matrix obtained from $U$ by replacing every nonzero singular value $i$ in $U$ by $\frac{1}{i}$, leaving all zeros in place, and then transposing.
Ridge Regression: Least-Squares Solutions

One solution is to pick the unique vector $w^+$ of smallest Euclidean norm $\|w^+\|_2$ that minimizes $\|Aw - y\|_2^2$.

The solution $w^+$ is given by $w^+ = A^+ y$, where $A^+$ is the pseudo-inverse of $A$. 
Ridge Regression: Least-Squares Solutions

One solution is to pick the unique vector $w^+$ of smallest Euclidean norm $\|w^+\|_2$ that minimizes $\|Aw - y\|^2_2$.

The solution $w^+$ is given by $w^+ = A^+ y$, where $A^+$ is the pseudo-inverse of $A$.

The matrix $A^+$ is obtained from an SVD of $A$, say $A = V\Sigma U^\intercal$. 
**Ridge Regression: Least-Squares Solutions**

One solution is to pick the unique vector $w^+$ of smallest Euclidean norm $\|w^+\|_2$ that minimizes $\|Aw - y\|_2^2$.

The solution $w^+$ is given by $w^+ = A^+y$, where $A^+$ is the pseudo-inverse of $A$.

The matrix $A^+$ is obtained from an SVD of $A$, say $A = V\Sigma U^\top$.

Namely, $A^+ = U\Sigma^+ V^\top$, where $\Sigma^+$ is the matrix obtained from $\Sigma$ by replacing every nonzero singular value $\sigma_i$ in $\Sigma$ by $\sigma_i^{-1}$, leaving all zeros in place, and then transposing.
Ridge Regression: Regularization Term

The difficulty with this approach is that it requires knowing whether a singular value is zero or very small but nonzero.
Ridge Regression: Regularization Term

The difficulty with this approach is that it requires knowing whether a singular value is zero or very small but nonzero.

A very small nonzero singular value $\sigma$ in $\Sigma$ yields a very large value $\sigma^{-1}$ in $\Sigma^+$, but $\sigma = 0$ remains 0 in $\Sigma^+$. 
Ridge Regression: Regularization Term

The difficulty with this approach is that it requires knowing whether a singular value is zero or very small but nonzero.

A very small nonzero singular value $\sigma$ in $\Sigma$ yields a very large value $\sigma^{-1}$ in $\Sigma^+$, but $\sigma = 0$ remains 0 in $\Sigma^+$.  

This discontiuity phenomenon is not desirable and another way is to control the size of $w$ by adding a regularization term to $\|Aw - y\|^2$, and a natural candidate is $\|w\|^2$.
Ridge Regression: Notational Convention

It is customary to rename each column vector $a_i \top$ as $x_i$ (where $x_i \in \mathbb{R}^n$) and to rename the input data matrix $A$ as $X$, so that the row vector $x_i \top$ are the rows of the $m \times n$ matrix $X$

$$X = \begin{pmatrix} x_1 \top \\ \vdots \\ x_m \top \end{pmatrix}.$$
Ridge Regression: Program (RR1)

Our optimization problem, called ridge regression, is
Ridge Regression: Program (RR1)

Our optimization problem, called *ridge regression*, is

Program (RR1):

\[
\text{minimize} \quad \|y - Xw\|^2 + K\|w\|^2,
\]
Ridge Regression: Program (RR1)

Our optimization problem, called *ridge regression*, is

**Program (RR1):**

\[
\text{minimize} \quad \|y - Xw\|^2 + K\|w\|^2,
\]

which by introducing the new variable \(\xi = y - Xw\) can be rewritten as
Ridge Regression: Program (RR2)

Program (RR2):

minimize $\xi^T \xi + Kw^T w$
subject to $y - Xw = \xi$,

where $K > 0$ is some constant determining the influence of the regularizing term $w^T w$, and we minimize over $\xi$ and $w$. 
Ridge Regression: Program (RR1) Solution

The objective function of the first version of our minimization problem can be expressed as

\[
J(w) = \|y - Xw\|^2 + K\|w\|^2
\]

\[
= w^T (X^T X + Kl_n)w - 2w^T X^T y + y^T y.
\]
**Ridge Regression: Program (RR1) Solution**

The objective function of the first version of our minimization problem can be expressed as

\[ J(w) = \| y - Xw \|^2 + K \| w \|^2 \]

\[ = w^T (X^T X + KI_n)w - 2w^T X^T y + y^T y. \]

The matrix \( X^T X \) is symmetric positive semidefinite and \( K > 0 \), so the matrix \( X^T X + KI_n \) is *positive definite.*
Ridge Regression: Program (RR1) Solution

The objective function of the first version of our minimization problem can be expressed as

\[ J(w) = \|y - Xw\|^2 + K\|w\|^2 = w^T (X^T X + Kl_n)w - 2w^T X^T y + y^T y. \]

The matrix \( X^T X \) is symmetric positive semidefinite and \( K > 0 \), so the matrix \( X^T X + Kl_n \) is positive definite.

It follows that \( J \) is strictly convex, so by a previous theorem it has a unique minimum iff \( \nabla J_w = 0 \).
Ridge Regression: Program (RR1) Solution

Since
\[ \nabla J_w = 2(X^\top X + Kl_n)w - 2X^\top y, \]
we deduce that
\[ w = (X^\top X + Kl_n)^{-1}X^\top y. \] (*_{wp})
Ridge Regression: Program (RR1) Solution

Since
\[ \nabla J_w = 2(X^\top X + Kl_n)w - 2X^\top y, \]
we deduce that
\[ w = (X^\top X + Kl_n)^{-1}X^\top y. \] (*wp)

There is an interesting connection between the matrix \((X^\top X + Kl_n)^{-1}X^\top\) and the pseudo-inverse \(X^+\) of \(X\).
Ridge Regression: Program (RR1) Solution

Since
\[
\nabla J_w = 2(X^\top X + Kl_n)w - 2X^\top y,
\]
we deduce that
\[
w = (X^\top X + Kl_n)^{-1}X^\top y. \tag{*_{wp}}
\]

There is an interesting connection between the matrix \((X^\top X + Kl_n)^{-1}X^\top\) and the pseudo-inverse \(X^+\) of \(X\).

**Proposition.** The limit of the matrix \((X^\top X + Kl_n)^{-1}X^\top\) when \(K > 0\) goes to zero is the pseudo-inverse \(X^+\) of \(X\).
Ridge Regression: Program (RR2) Solution

The dual function of the first formulation of our problem is a constant function (with value the minimum of $J$) so it is not useful, but the second formulation of our problem yields an interesting dual problem.
Ridge Regression: Program (RR2) Solution

The dual function of the first formulation of our problem is a constant function (with value the minimum of $J$) so it is not useful, but the second formulation of our problem yields an interesting dual problem.

The Lagrangian is

$$L(\xi, w, \lambda) = \xi^\top \xi + Kw^\top w + (y - Xw - \xi)^\top \lambda$$

$$= \xi^\top \xi + Kw^\top w - w^\top X^\top \lambda - \xi^\top \lambda + \lambda^\top y,$$

with $\lambda, \xi, y \in \mathbb{R}^m$. 
Ridge Regression: Program (RR2) Solution

The dual function of the first formulation of our problem is a constant function (with value the minimum of $J$) so it is not useful, but the second formulation of our problem yields an interesting dual problem.

The Lagrangian is

$$L(\xi, w, \lambda) = \xi^T \xi + Kw^T w + (y - Xw - \xi)^T \lambda$$

$$= \xi^T \xi + Kw^T w - w^T X^T \lambda - \xi^T \lambda + \lambda^T y,$$

with $\lambda, \xi, y \in \mathbb{R}^m$.

The Lagrangian $L(\xi, w, \lambda)$, as a function of $\xi$ and $w$ with $\lambda$ held fixed, is obviously convex, in fact strictly convex.
**Ridge Regression: Dual Function of (RR2)**

To derive the dual function $G(\lambda)$ we minimize $L(\xi, w, \lambda)$ with respect to $\xi$ and $w$. 
Ridge Regression: Dual Function of \((RR2)\)

To derive the dual function \(G(\lambda)\) we minimize \(L(\xi, w, \lambda)\) with respect to \(\xi\) and \(w\).

Since \(L(\xi, w, \lambda)\) is (strictly) convex as a function of \(\xi\) and \(w\), by a previous theorem it has a minimum iff its gradient \(\nabla L_{\xi, w}\) is zero.
Ridge Regression: Dual Function of \((RR2)\)

Since

\[
\nabla L_{\xi, w} = \begin{pmatrix}
2\xi - \lambda \\
2 Kw - X^\top \lambda
\end{pmatrix},
\]
Ridge Regression: Dual Function of (RR2)

Since

\[ \nabla L_{\xi,w} = \begin{pmatrix} 2\xi - \lambda \\ 2Kw - X^T \lambda \end{pmatrix}, \]

we get

\[ \lambda = 2\xi \]

\[ w = \frac{1}{2K}X^T \lambda = X^T \frac{\xi}{K}. \]
Ridge Regression: Dual Function of (RR2)

The above suggests defining the variable $\alpha$ so that $\xi = K\alpha$, so we have $\lambda = 2K\alpha$ and $w = X^\top\alpha$. 
Ridge Regression: Dual Function of (RR2)

The above suggests defining the variable $\alpha$ so that $\xi = K\alpha$, so we have $\lambda = 2K\alpha$ and $w = X^\top\alpha$.

Then we obtain the dual function as a function of $\alpha$ by substituting the above values of $\xi$, $\lambda$ and $w$ back in the Lagrangian, and we get

$$G(\alpha) = -K\alpha^\top(XX^\top + KI_m)\alpha + 2K\alpha^\top y.$$
Ridge Regression: Problem (RR2) Solution

This is a strictly concave function so by a previous theorem its maximum is achieved iff $\nabla G_\alpha = 0$, that is,

$$2K(XX^\top + KI_m)\alpha = 2Ky,$$

which yields

$$\alpha = (XX^\top + KI_m)^{-1}y.$$
Ridge Regression: Solution Comparison

Putting everything together we obtain

\[ \alpha = (XX^T + Kl_m)^{-1} y \]

\[ w = X^\top \alpha \]

\[ \xi = K\alpha, \]

which yields

\[ w = X^\top (XX^T + Kl_m)^{-1} y. \]
Ridge Regression

Earlier in (*_wp) we found that

\[ w = (X^T X + KI_n)^{-1} X^T y, \]

and it is easy to check that

\[ (X^T X + KI_n)^{-1} X^T = X^T (XX^T + KI_m)^{-1}. \]
Ridge Regression

Earlier in \(*_{wp}\) we found that

\[
    w = (X^\top X + Kl_n)^{-1}X^\top y,
\]

and it is easy to check that

\[
    (X^\top X + Kl_n)^{-1}X^\top = X^\top (XX^\top + Kl_m)^{-1}.
\]

If \(n < m\) it is cheaper to use the formula on the left-hand side, but if \(m < n\) it is cheaper to use the formula on the right-hand side.