Optimization
MS Maths Big Data

Alexandre Gramfort

alexandre.gramfort@telecom-paristech.fr

Telecom ParisTech

M2 Maths Big Data

Plan

Optimization problem

Definition (Optimization problem ($\mathcal{P}$))

- $\min f(x)$, $x \in \mathcal{C}$, where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called the **objective function**
- $\mathcal{C} = \{x \in \mathbb{R}^n / g(x) \leq 0 \text{ et } h(x) = 0\}$ is the **feasible set**
- $g(x) \leq 0$ represent **inequality constraints**.
  $g(x) = (g_1(x), \ldots, g_p(x))$ so with p contraints.
- $h(x) = 0$ represent **equality contraints**.
  $h(x) = (h_1(x), \ldots, h_q(x))$ so with q contraints.
- an element $x \in \mathcal{C}$ is said to be **feasible**

Taylor at order 2

Assuming f is twice differentiable, the Taylor expansion at order 2 of f at x reads:

$$\forall h \in \mathbb{R}^n,\ f(x+h) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^T \nabla^2 f(x) h + o(\|h\|^2)$$

- $\nabla f(x) \in \mathbb{R}^n$ is the gradient.
- $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ the Hessian matrix.

Remark: Local quadratic approximation

Plan

Ridge regression

We consider problems with n samples, observations, and p features, variables.

Definition (Ridge regression)

Let $y \in \mathbb{R}^n$ the n targets to predict and $(x_i)_i$ the n samples in $\mathbb{R}^p$. Ridge regression consists in solving the following problem

$$\min_{w,b} \frac{1}{2}\|y - Xw - b\|^2 + \frac{\lambda}{2}\|w\|^2, \lambda > 0$$

where $w \in \mathbb{R}^p$ is called the weights vector, $b \in \mathbb{R}$ is the intercept (a.k.a. bias) and the ith row of X is x_i.

Remark: : Note that the intercept is not penalized with λ.

Taking care of the intercept

There are different ways to deal with the intercept.

- Option 1: Center the target y and each column feature. After centering the problem reads:

$$\min_w \frac{1}{2}\|y - Xw\|^2 + \frac{\lambda}{2}\|w\|^2, \lambda > 0$$

- Option 2: Add a column of 1 to X and try not to penalize it (too much).

Exercise

- Denote by $\overline{y} \in \mathbb{R}$ the mean of y and by $\overline{X} \in \mathbb{R}^p$ the mean of each column of X. Show that $\hat{b} = -\overline{X}^T \hat{w} + \overline{y}$.

Ridge regression

Definition (Quadratic form)

A quadratic form reads

$$f(x) = \frac{1}{2} x^T A x + b^T x + c$$

where $x \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$, $b \in \mathbb{R}^p$ and $c \in \mathbb{R}$.

Ridge regression

Questions

- Show that ridge regression boils down to the minimization of a quadratic form.
- Propose a closed form solution.
- Show that the solution is obtained by solving a linear system.
- Is the objective function strongly convex?
- Assuming $n < p$ what is the value of the constant of strong convexity?

Plan

Singular value decomposition (SVD)

- SVD is a factorization of a matrix (real here)
- $M = U\Sigma V^T$ where $M \in \mathbb{R}^{n \times p}$, $U \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{n \times p}$, $V \in \mathbb{R}^{p \times p}$
- $U^T U = UU^T = I_n$ (orthogonal matrix)
- $V^T V = VV^T = I_p$ (orthogonal matrix)
- Σ diagonal matrix
- $\Sigma_{i,i}$ are called the singular values
- U are left-singular vectors
- V are right-singular vectors

Singular value decomposition (SVD)

- SVD is a factorization of a matrix (real here)
- U contains the eigenvectors of MM^T associated to the eigenvalues $\Sigma_{i,i}^2$
- V contains the eigenvectors of $M^T M$ associated to the eigenvalues $\Sigma_{i,i}^2$
- we assume here $\Sigma_{i,i} = 0$ for $\min(n, p) \leq i \leq \max(n, p)$
- SVD is particularly useful to find the rank, null-space, image and pseudo-inverse of a matrix

Plan

Matrix inversion lemma

Proposition (Matrix inversion lemma)

also known as Sherman–Morrison–Woodbury formula states that:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1}VA^{-1},$$

where $A \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times k}$, $V \in \mathbb{R}^{k\times n}$.

Matrix inversion lemma (proof)

Just check that (A+UCV) times the RHS of the Woodbury identity gives the identity matrix:

$$
\begin{aligned}
&(A + UCV)\left[A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1}VA^{-1}\right] \\
&\quad = I + UCVA^{-1} - (U + UCVA^{-1}U)(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \\
&\quad = I + UCVA^{-1} - UC(C^{-1} + VA^{-1}U)(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \\
&\quad = I + UCVA^{-1} - UCVA^{-1} = I
\end{aligned}
$$

Questions

- Using the matrix inversion lemma show that if $n < p$, the ridge regression problem can be solved by inverting a matrix of size $n \times n$ rather than $p \times p$.

Plan

Primal and dual implementation

The solution of the ridge regression problem (without intercept) is obtained by solving the problem in the primal form:

$$\hat{w} = (X^T X + \lambda I_p)^{-1} X^T y$$

or in the dual form:

$$\hat{w} = X^T (X X^T + \lambda I_n)^{-1} y$$

In the dual formulation the matrix to invert in $\mathbb{R}^{n \times n}$.

What if X is sparse, n is 1e5 and p is 1e6?

Primal and dual implementation

The solution of the ridge regression problem (without intercept) is obtained by solving the problem in the primal form:

$$\hat{w} = (X^T X + \lambda I_p)^{-1} X^T y$$

or in the dual form:

$$\hat{w} = X^T (X X^T + \lambda I_n)^{-1} y$$

In the dual formulation the matrix to invert in $\mathbb{R}^{n \times n}$.

What if X is sparse, n is 1e5 and p is 1e6?

Plan

Conjugate gradient: Solve $Ax = b$, $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$

$x_0 \in \mathbb{R}^n$, $g_0 = Ax_0 - b$
for $k = 0$ to n **do**
  if $g_k = 0$ **then**
    break
  end if
  if $k = 0$ **then**
    $w_k = g_0$
  else
    $\alpha_k = -\frac{\langle g_k, Aw_{k-1} \rangle}{\langle w_{k-1}, Aw_{k-1} \rangle}$
    $w_k = g_k + \alpha_k w_{k-1}$
  end if
  $\rho_k = \frac{\langle g_k, w_k \rangle}{\langle w_k, Aw_k \rangle}$
  $x_{k+1} = x_k - \rho_k w_k$
  $g_{k+1} = Ax_{k+1} - b$
end for
return x_{k+1}

Sparse ridge with CG

cf. Notebook

Logistic regression with CG

cf. Notebook

Warm starts and paths

In machine learning it is common to try to solve a problem that is very similar to a previous one.

- You train a model every day and you need just to "update" the model
- You look for the best hyperparmater and evaluate the parameter on a grid of values. For example on a grid of λ.