Recitation 11

Carles Domingo

Least squares

Remember: The least squares problems can be written as

$$\min_{x \in \mathbb{R}^d} \|Ax - y\|^2, \tag{1}$$

where $A \in \mathbb{R}^{n \times d}$. And by the first order condition for minimizers of convex functions,

$$x \text{ is a solution of (1)} \iff A^\top A x = A^\top y$$

Definition (Moore-Penrose pseudo-inverse)

If $A = U\Sigma V^\top$, then $A^\dagger = V\Sigma' U^\top \in \mathbb{R}^{d \times n}$ is the Moore-Penrose pseudo-inverse of A, where $\Sigma' \in \mathbb{R}^{d \times n}$ is defined as

$$\Sigma'_{ii} = \begin{cases} 1/\Sigma_{ii} & \text{when } \Sigma_{ii} \neq 0 \\ 0 & \text{otherwise} \end{cases}, \quad \Sigma'_{ij} = 0 \text{ when } i \neq j$$

Least squares

Theorem (Unregularized least squares)

The set of solutions of the minimization problem $\min_{x\in\mathbb{R}^d}\|Ax-y\|^2$ *is* $A^\dagger y + \mathit{Ker}(A)$.

Theorem (Ridge regression)

For any $\lambda > 0$, *the unique solution of the minimization problem* $\min_{x\in\mathbb{R}^d}\{\|Ax-y\|^2+\lambda\|x\|^2\}$ *is*

$$x^{\mathit{ridge}} = (A^\top A + \lambda \mathit{Id})^{-1}A^\top y$$

Definition (Lasso)

The Lasso x^{Lasso} is defined as
$x^{\text{Lasso}} = \arg\min_{x\in\mathbb{R}^d}\{\|Ax-y\|^2+\lambda\|x\|_1\}$.

Ridge regression

Show that the solution x^{ridge} of ridge regression is given by the formula in the previous slide, i.e.

$$x^{\text{ridge}} = (A^\top A + \lambda \mathsf{Id})^{-1} A^\top y$$

Ridge regression

Show that the solution x^{ridge} of ridge regression is given by the formula in the previous slide, i.e.

$$x^{\text{ridge}} = (A^\top A + \lambda \mathsf{Id})^{-1} A^\top y$$

Ridge regression

Show that the solution x^{ridge} of ridge regression is given by the formula in the previous slide, i.e.

$$x^{\text{ridge}} = (A^\top A + \lambda \mathsf{Id})^{-1} A^\top y$$

Lasso for orthogonal A

1. For $x_0 \in \mathbb{R}$, let $f_{x_0} : \mathbb{R} \to \mathbb{R}$ be defined as $f_{x_0}(x) = \frac{1}{2}x^2 - x_0 x + \lambda|x|$. Show that for $\lambda \geq 0$, the function f_{x_0} admits a unique minimizer given by $x^* = \eta(x_0; \lambda)$, where η is the *soft-thresholding* function:

$$\eta(x_0; \lambda) = \begin{cases} x_0 - \lambda & \text{if } x_0 \geq \lambda \\ 0 & \text{if } -\lambda \leq x_0 \leq \lambda \\ x_0 + \lambda & \text{if } x_0 \leq -\lambda \end{cases}$$

Lasso for orthogonal A

1. For $x_0 \in \mathbb{R}$, let $f_{x_0} : \mathbb{R} \to \mathbb{R}$ be defined as $f_{x_0}(x) = \frac{1}{2}x^2 - x_0 x + \lambda|x|$. Show that for $\lambda \geq 0$, the function f_{x_0} admits a unique minimizer given by $x^* = \eta(x_0; \lambda)$, where η is the *soft-thresholding* function:

$$\eta(x_0; \lambda) = \begin{cases} x_0 - \lambda & \text{if } x_0 \geq \lambda \\ 0 & \text{if } -\lambda \leq x_0 \leq \lambda \\ x_0 + \lambda & \text{if } x_0 \leq -\lambda \end{cases}$$

Lasso for orthogonal A

1. For $x_0 \in \mathbb{R}$, let $f_{x_0} : \mathbb{R} \to \mathbb{R}$ be defined as $f_{x_0}(x) = \frac{1}{2}x^2 - x_0 x + \lambda|x|$. Show that for $\lambda \geq 0$, the function f_{x_0} admits a unique minimizer given by $x^* = \eta(x_0; \lambda)$, where η is the *soft-thresholding* function:

$$\eta(x_0; \lambda) = \begin{cases} x_0 - \lambda & \text{if } x_0 \geq \lambda \\ 0 & \text{if } -\lambda \leq x_0 \leq \lambda \\ x_0 + \lambda & \text{if } x_0 \leq -\lambda \end{cases}$$

Lasso for orthogonal A

1. For $x_0 \in \mathbb{R}$, let $f_{x_0} : \mathbb{R} \to \mathbb{R}$ be defined as $f_{x_0}(x) = \frac{1}{2}x^2 - x_0 x + \lambda|x|$. Show that for $\lambda \geq 0$, the function f_{x_0} admits a unique minimizer given by $x^* = \eta(x_0; \lambda)$, where η is the *soft-thresholding* function:

$$\eta(x_0; \lambda) = \begin{cases} x_0 - \lambda & \text{if } x_0 \geq \lambda \\ 0 & \text{if } -\lambda \leq x_0 \leq \lambda \\ x_0 + \lambda & \text{if } x_0 \leq -\lambda \end{cases}$$

Lasso for orthogonal A

2. Let $A \in \mathbb{R}^{n\times d}$ be a matrix such that its columns are orthonormal (i.e. $A^\top A = \mathrm{Id}$). Show that the Lasso solution $x^{\mathsf{Lasso}} = \arg\min_{x\in\mathbb{R}^d}\{\|Ax-y\|^2 + \lambda\|x\|_1\}$ satisfies

$$x_j^{\mathsf{Lasso}} = \eta(x_j^{\mathsf{LS}};\lambda), \quad \forall j \in 1,\ldots,d,$$

where $x^{\mathsf{LS}} = A^\dagger y$.

Lasso for orthogonal A

2. Let $A \in \mathbb{R}^{n\times d}$ be a matrix such that its columns are orthonormal (i.e. $A^\top A = \mathrm{Id}$). Show that the Lasso solution $x^{\text{Lasso}} = \arg\min_{x\in\mathbb{R}^d}\{\|Ax-y\|^2 + \lambda\|x\|_1\}$ satisfies

$$x_j^{\text{Lasso}} = \eta(x_j^{\text{LS}};\lambda), \quad \forall j \in 1,\dots,d,$$

where $x^{\text{LS}} = A^\dagger y$.

Moore-Penrose pseudo-inverse

In this exercise we will show that for $A \in \mathbb{R}^{n\times d}$, the Moore-Penrose pseudo-inverse $A^\dagger \in \mathbb{R}^{d\times n}$ of A is the only matrix in $\mathbb{R}^{d\times n}$ such that

1. $AA^\dagger A = A$.
2. $A^\dagger A A^\dagger = A^\dagger$.
3. $AA^\dagger \in \mathbb{R}^{n\times n}$ and $A^\dagger A \in \mathbb{R}^{d\times n}$ are symmetric matrices.

We do this in two steps:

1. Show that the Moore-Penrose pseudo-inverse as defined in the second slide fulfills (1), (2), (3).
2. Show that for a given $A \in \mathbb{R}^{n\times d}$, there exists a unique matrix $A^\dagger \in \mathbb{R}^{d\times n}$ fulfilling (1), (2), (3).

Moore-Penrose pseudo-inverse

In this exercise we will show that for $A \in \mathbb{R}^{n\times d}$, the Moore-Penrose pseudo-inverse $A^\dagger \in \mathbb{R}^{d\times n}$ of A is the only matrix in $\mathbb{R}^{d\times n}$ such that

1. $AA^\dagger A = A$.
2. $A^\dagger AA^\dagger = A^\dagger$.
3. $AA^\dagger \in \mathbb{R}^{n\times n}$ and $A^\dagger A \in \mathbb{R}^{d\times n}$ are symmetric matrices.

We do this in two steps:

1. Show that the Moore-Penrose pseudo-inverse as defined in the second slide fulfills (1), (2), (3).
2. Show that for a given $A \in \mathbb{R}^{n\times d}$, there exists a unique matrix $A^\dagger \in \mathbb{R}^{d\times n}$ fulfilling (1), (2), (3).

Moore-Penrose pseudo-inverse

1. $AA^\dagger A = A$.
2. $A^\dagger A A^\dagger = A^\dagger$.
3. $AA^\dagger \in \mathbb{R}^{n\times n}$ and $A^\dagger A \in \mathbb{R}^{d\times n}$ are symmetric matrices.

Show that the Moore-Penrose pseudo-inverse as defined in the second slide fulfills (1), (2), (3).

Moore-Penrose pseudo-inverse

1. $AA^\dagger A = A$.
2. $A^\dagger A A^\dagger = A^\dagger$.
3. $AA^\dagger \in \mathbb{R}^{n\times n}$ and $A^\dagger A \in \mathbb{R}^{d\times n}$ are symmetric matrices.

Show that the Moore-Penrose pseudo-inverse as defined in the second slide fulfills (1), (2), (3).

Moore-Penrose pseudo-inverse

1. $AA^\dagger A = A$.
2. $A^\dagger A A^\dagger = A^\dagger$.
3. $AA^\dagger \in \mathbb{R}^{n\times n}$ and $A^\dagger A \in \mathbb{R}^{d\times n}$ are symmetric matrices.

Show that the Moore-Penrose pseudo-inverse as defined in the second slide fulfills (1), (2), (3).

Moore-Penrose pseudo-inverse

1. $AA^\dagger A = A$.
2. $A^\dagger AA^\dagger = A^\dagger$.
3. $AA^\dagger \in \mathbb{R}^{n\times n}$ and $A^\dagger A \in \mathbb{R}^{d\times n}$ are symmetric matrices.

Show that for a given $A \in \mathbb{R}^{n\times d}$, there exists a unique matrix $A^\dagger \in \mathbb{R}^{d\times n}$ fulfilling (1), (2), (3).

Moore-Penrose pseudo-inverse

1. $AA^\dagger A = A$.
2. $A^\dagger A A^\dagger = A^\dagger$.
3. $AA^\dagger \in \mathbb{R}^{n\times n}$ and $A^\dagger A \in \mathbb{R}^{d\times n}$ are symmetric matrices.

Show that for a given $A \in \mathbb{R}^{n\times d}$, there exists a unique matrix $A^\dagger \in \mathbb{R}^{d\times n}$ fulfilling (1), (2), (3).

Moore-Penrose pseudo-inverse

1. $AA^\dagger A = A$.
2. $A^\dagger A A^\dagger = A^\dagger$.
3. $AA^\dagger \in \mathbb{R}^{n\times n}$ and $A^\dagger A \in \mathbb{R}^{d\times n}$ are symmetric matrices.

Show that for a given $A \in \mathbb{R}^{n\times d}$, there exists a unique matrix $A^\dagger \in \mathbb{R}^{d\times n}$ fulfilling (1), (2), (3).

Extra: Bayesian view

This exercise studies a simple setting in Bayesian statistics where ridge regression appears. Suppose that we have n data points $(a_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ and we know that for all $i \in [1 : n]$, $y_i = x^\top a_i + \epsilon_i$, where ϵ_i has a standard normal distribution and x has a prior $\mathcal{N}(0, \sigma^2 Id)$. By Bayes' theorem, the posterior density for x is

$$\mathbb{P}(x \mid (a_i, y_i)_{i=1}^n) = \frac{\mathbb{P}((a_i, y_i)_{i=1}^n \mid x)\mathbb{P}(x)}{\mathbb{P}((a_i, y_i)_{i=1}^n)}$$

The maximum a posteriori (MAP) estimator is defined as

$$\hat{x} = \arg\max_{x \in \mathbb{R}^d} \mathbb{P}((a_i, y_i)_{i=1}^n \mid x)\mathbb{P}(x).$$

Show that the MAP estimator computation corresponds to solving a ridge regression problem.