02471 Machine Learning for Signal Processing

Week 6: Sparsity-aware learning with ℓ_1

This exercise is based on S. Theodoridis: Machine Learning, A Bayesian and Optimization Perspective 2nd edition, section(s) 9.1–9.5, 9.9.

The objective of this exercise is to get better acquainted with learning sparse solutions using ℓ_1 regularization and touch upon compressed sensing.

Overview

The exercise have the following structure:

6.1 will prove important results we need for norms.

6.2 will manually calculate the solution for least squares using different norms as regularization.

6.3 will use the ℓ_1 norm to carry out compressed sensing and signal recovery for a toy example.

6.4 will demonstrate a use-case where we use compressed sensing to recover a multi-tone signal using fewer samples than usually required.

Notation

- $J(\boldsymbol{\theta})$ is a cost function that we are seeking to minimize with respect to $\boldsymbol{\theta}$.
- n indicates the time step as we collect data $(y_n, \boldsymbol{x}_n)$, where y_n is our target, or desired signal, and $\boldsymbol{x}_n$ is our filter input.
- $\mathcal{N}(\mu, \sigma^2)$ denotes a normal (Gaussian) distribution with mean value μ and variance σ^2.

Code

The code can be found in the .m and .py files named in the same way as exercises, ie. the code for exercise 6.x.y is in the file 6_x_y.m (or .py).

For coding exercises that requires implementation we will usually write `complete this line` where the implementing should be done.

Solutions

The solution is provided for all derivation exercises, and often hints are provided at the end of the document. If you get stuck, take a look at the hints, and if you are still stuck, take a look in the solution to see the approach being taken. Then try to do it on your own.

Solutions are also provided for some coding exercises. If you get stuck, take a look at the solution, and then try to implement it on your own.

6.1 Norms

In this exercise we will work with norms. Be sure to read 9.2 in the book before carrying out this exercise.

A function, $\|\cdot\|_p : \mathbb{R}^l \longmapsto [0, \infty)$ is a valid norm if the following properties hold:

1. $\|\boldsymbol{\theta}\|_p \geq 0, \quad \|\boldsymbol{\theta}\|_p = 0 \Leftrightarrow \boldsymbol{\theta} = \mathbf{0}$.
2. $\|\alpha\boldsymbol{\theta}\|_p = |\alpha|\|\boldsymbol{\theta}\|_p, \forall \alpha \in \mathbb{R}$.
3. $\|\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\|_p \leq \|\boldsymbol{\theta}_1\|_p + \|\boldsymbol{\theta}_2\|_p$ (triangle inequality).

We will work with the ℓ_p norm, which for $p =]0; \infty[$ is defined as

$$\|\theta\|_p := \left(\sum_{i=1}^{l} |\theta_i|^p\right)^{1/p}$$

For $p \to 0$ we have

$$\|\theta\|_0 := \sum_{i=1}^{l} \chi_{(0,\infty)}(|\theta_i|)$$

$$\chi_{\mathcal{A}}(\tau) := \begin{cases} 1, & \text{if } \tau \in \mathcal{A} \\ 0, & \text{if } \tau \notin \mathcal{A} \end{cases}$$

Exercise 6.1.1

In the search for the sparsest solution, we want to use the $\|\cdot\|_0$ norm. However, this is not a true norm unfortunately. Show that $\|\cdot\|_p$ for $0 \leq p < 1$ is not a true norm.

6.2 The regularized least-squares solution

In this exercise we will work with the norm shrinkage solution. Be sure to skim sec. 9.3 before carrying out this exercise.

We will work with the regression task

$$\boldsymbol{y} = X\boldsymbol{\theta} + \boldsymbol{\eta}, \quad \boldsymbol{y} \in \mathbb{R}^N, \, X \in \mathbb{R}^{N \times l}, \, \boldsymbol{\theta} \in \mathbb{R}^l, \, \boldsymbol{\eta} \in \mathbb{R}^N,$$

where we assume $\boldsymbol{\eta}$ is generated using a white noise sequence.

We will estimate the solution using the cost function

$$J(\boldsymbol{\theta}) = \|\boldsymbol{y} - X\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_\ell^2$$

where, for $\lambda = 0$, we obtain the least squares estimate, for $\lambda > 0$ and $\ell = 2$ we obtain Ridge regression, and for $\lambda > 0$ and $\ell = 1$ we obtain LASSO regression. If we assume the regressors have the property $X^T X = I$, the solutions to our regression problem are:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{LS} &= X^T\boldsymbol{y} \\ \hat{\boldsymbol{\theta}}_R &= \frac{1}{1+\lambda}\hat{\boldsymbol{\theta}}_{LS} \\ \hat{\theta}_{1,i} &= \operatorname{sgn}\left(\hat{\theta}_{LS,i}\right)\left(|\hat{\theta}_{LS,i}| - \frac{\lambda}{2}\right)_+ \qquad i = 1, 2, \cdots, l \end{aligned}$$

Exercise 6.2.1

The solutions for least squares regression and Ridge regression are

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \left(X^T X\right)^{-1} X^T\boldsymbol{y} = X^T\boldsymbol{y}$$

$$\hat{\boldsymbol{\theta}}_R = \left(X^T X + \lambda I\right)^{-1} X^T\boldsymbol{y}$$

Prove the result $\hat{\boldsymbol{\theta}}_R = \frac{1}{1+\lambda}\hat{\boldsymbol{\theta}}_{LS}$.

Exercise 6.2.2

For LASSO regression, taking the gradient of the cost function and equating to zero now yields a set (there are multiple solutions)

$$\mathbf{0} \in -2X^T\boldsymbol{y} + 2X^T X\boldsymbol{\theta} + \lambda\partial\|\boldsymbol{\theta}\|_1$$

If $X^T X = I$, we get

$$0 \in -\hat{\theta}_{\mathrm{LS},i} + \hat{\theta}_{1,i} + \frac{\lambda}{2}\partial\left|\hat{\theta}_{1,i}\right|, \quad \forall i$$

$$\partial|\theta| = \begin{cases} \{1\}, & \text{if } \theta > 0 \\ \{-1\}, & \text{if } \theta < 0 \\ [-1, 1], & \text{if } \theta = 0 \end{cases}$$

Based on the information presented to you, show that

$$\hat{\theta}_{1,i} = \operatorname{sgn}\left(\hat{\theta}_{LS,i}\right)\left(|\hat{\theta}_{LS,i}| - \frac{\lambda}{2}\right)_+ \qquad i = 1, 2, \cdots, l$$

Exercise 6.2.3

Assume now we have a least-squares solution vector $\hat{\boldsymbol{\theta}}_{LS} = [0.7, -0.3, 0.1, -2]^T$. Compute the corresponding Ridge and LASSO solution for $\lambda = 1$.

Exercise 6.2.4

For what value of λ will Ridge regression result in the null vector? For what value of λ will LASSO regression result in the null vector?

6.3 Reconstruct a sparse vector using compressed sensing

In this exercise we will replicate example 9.5 and reconstruct a sparse vector $\boldsymbol{\theta} \in \mathbb{R}^l$, with its first components taking random values drawn from a normal distribution, $\mathcal{N}(0, 1)$ and the rest being equal to zero.

To recover $\boldsymbol{\theta}$, we build a sensing matrix X with N rows having samples normally distributed as $\mathcal{N}(0, 1/N)$. Then we solve the regression problem using LASSO and Ridge regression and compare the results and recovery rate.

Exercise 6.3.1

Inspect and complete the code associated with this exercise. What is the parameters λ, N, and how many non-zeros does $\boldsymbol{\theta}$ have? What is the figure showing? Experiment with different parameters – can you make the ridge regression have equal performance to the LASSO?

Exercise 6.3.2

Inspect and complete the code associated with this exercise. The code repeats the previous experiment and counts the number of successful recoveries. How is this calculated? Use the number to estimate the recovery probability. What is the figure showing? Experiment with the different parameters, e.g. how many observations do you need to have a high recovery probability?

6.4 Signal recovery of a multi-tone signal

In this exercise we will build upon the previous exercise and construct a multi-tone signal and recover the signal in a compressed sensing situation. This could be e.g that the signal is transmitted through a communication channel, and now we need to estimate the tones that was transmitted.

The multi-tone signal is sampled as

$$x_n = \sum_{j=1}^{3} a_j \cos\Big(\frac{\pi}{2l}(2m_j - 1)n\Big), \quad n = 0, ..., l-1,$$

where $N = 30$, $l = 2^8$, $\boldsymbol{a} = [0.3, 1, 0.75]^T$ and $\boldsymbol{m} = [4, 10, 30]^T$.

Exercise 6.4.1

Inspect and complete the code associated with this exercise. What is the figure showing?

Exercise 6.4.2

To recover $\boldsymbol{x}$, we build a sensing matrix $X \in \mathbb{R}^{30\times 2^8}$ with entries drawn from a normal distribution $\mathcal{N}(0, 1/N)$ and use LASSO regression to recover the signal. This is not at all a realistic scenario in practice, and only used to demonstrate the theory works

Inspect and complete the code associated with this exercise.

What is the figure showing? Experiment with the different parameters.

Exercise 6.4.3

To make the situation more realistic, imagine now that we are doing random sampling of the signal (as opposed to uniform sampling with a sampling frequency abiding the nyquest rate.)

We will now use a sparse sensing matrix instead, where each row of X is a 1–sparse vector with the non-zero element being set to 1. Moreover, each column has at most one nonzero component. What operation will this sensing matrix carry out?

Inspect and complete the code associated with this exercise.

Empirically show (using the code) that $\boldsymbol{x}$ can be recovered exactly using such a sparse sensing matrix. Observe that the unknown $\boldsymbol{x}$ is sparse in the frequency domain and give an explanation why the recovery is successful with the specific sparse sensing matrix.

You may get different number of estimated components on each run. Complete the code to have a selection criteria for λ.

HINTS

Exercise 6.1.1

This is easiest proven by falsification, i.e. providing an example of vector(s) where the norm conditions does not hold.

Consider the cases $p = 0$ and $0 < p < 1$ separately.

Exercise 6.1.2

Property 1 and 2 are readily verified by substitution.

Exercise 6.2.1

Remember we assume $X^T X = I$.

Exercise 6.2.2

First try to arrive at the interim result by considering the different sign cases:

$$\hat{\theta}_{1,i} = \begin{cases} \hat{\theta}_{\mathrm{LS},i} - \frac{\lambda}{2}, & \text{if } \hat{\theta}_{1,i} > 0 \\ \hat{\theta}_{\mathrm{LS},i} + \frac{\lambda}{2}, & \text{if } \hat{\theta}_{1,i} < 0 \end{cases}$$

Exercise 6.2.3

Consider figure 9.3 in the book.