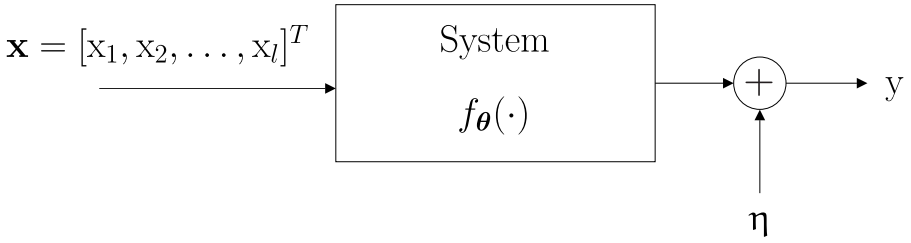
02471 Machine Learning for Signal Processing

Week 2: Parameter Estimation

This exercise is based on S. Theodoridis: Machine Learning, A Bayesian and Optimization Perspective 2nd edition, section(s) 3.1–3.3, 3.5, 3.8–3.11.

The objective of the exercise is to get acquainted with parameter estimation, the bias–variance dilemma and implications with key results.

We consider the following system $\mathrm{y} = f_{\boldsymbol{\theta}}(\mathbf{x}) + \eta$ (see figure).



For this particular exercise, we ignore the choice of $f_{\boldsymbol{\theta}}(\mathbf{x})$ and simply use linear regression models, and put focus on the estimation task, i.e. how to determine the parameters $\boldsymbol{\theta}$.

Overview

The exercise have the following structure:

2.1 is an exercise for linear regression, where we get acquainted with matrix manipulations and the bias-variance tradeoff.

2.2 will derive central results from ML section 3.5 and 3.9. This exercise will train you to get more comfortable with manipulating matrices and expectations, as well as show how some of the key results are derived.

2.3 is an exercise for ridge regression which is used to explore the bias-variance trade-off further.

Exercises can be carried out in any order, however it is beneficial to complete 2.1 first.

Notation

- Random variables are denoted with roman font, such as x, or θ, and their corresponding values (i.e. an observation) are denoted in italic, such as x or θ.
- Random vectors are denotes in bold roman font, such as $\mathbf{x}$ or $\boldsymbol{\theta}$, and their corresponding values are denoted in bold italic, such as $\boldsymbol{x}$ and $\boldsymbol{\theta}$.
- $J(\boldsymbol{\theta})$ is a cost function that we are seeking to minimize with respect to $\boldsymbol{\theta}$.
- θ denotes the parameters of a model. $\hat{\theta}$ denotes a point estimate of theta and θ_o denotes the optimal value for θ.

Code

The code can be found in the .m and .py files named in the same way as exercises, ie. the code for exercise 2.x.y is in the file 2_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.

2.1 Linear Models

To keep focus on the estimation task, we will use the simple linear regression model. The linear regression model, can for the n'th data point be written as

$$y_n = \theta_0 + \theta_1 x_{n,1} + \cdots + \theta_l n, l + \eta = \theta_0 + \boldsymbol{\theta}^T \mathbf{x} + \eta$$

Exercise 2.1.1

If you are unfamiliar with Linear Regression and Least Squares, work through example 3.1, page 74 in the ML book.

Exercise 2.1.2

The parameters of the linear regression model is estimated by minimizing the cost function

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}_b^T \boldsymbol{x}_n - \theta_0)^2$$

Let $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_b^T & \theta_0 \end{bmatrix}^T$ ($\boldsymbol{\theta}$ is a column vector). Define a matrix X and vector $\boldsymbol{y}$ such that the cost function can be rewritten to

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - X\boldsymbol{\theta})^T (\boldsymbol{y} - X\boldsymbol{\theta})$$

Validate the expressions are identical e.g. by using a 3-dimensional $\boldsymbol{\theta}$ vector and 2 datapoints. Make sure to familiarize yourself with these types of rewritings.

Exercise 2.1.3

Show that

$$\frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta}) = -2\mathrm{X}^T \boldsymbol{y} + 2\mathrm{X}^T X\boldsymbol{\theta}$$

We can now find the optimal set of parameters $\boldsymbol{\theta}$ by setting the derivative to zero, which then yields

$$(X^T X)\boldsymbol{\theta} = X^T \boldsymbol{y}$$

A few remarks to this solution.

- this is the prefered form for numerical solvers since these typically takes input on the form $A\boldsymbol{x} = \boldsymbol{b}$ where we solve for $\boldsymbol{x}$. We will use this later in the coding exercises.
- the cost function $J(\boldsymbol{\theta})$ is actually the squared error, and not the *mean* squared error (MSE), which is MSE $= \frac{1}{N}(\boldsymbol{y} - X\boldsymbol{\theta})^T(\boldsymbol{y} - X\boldsymbol{\theta})$. Note however, the factor $\frac{1}{N}$ vanishes when the derivative is set to zero, so the two functions have the same optimal solution.

Exercise 2.1.4

Inspect the code associated with this exercise (`ex2_1_4`). Run the code to create a training-set with a 2-dimensional input variable and a 1-dimensional output variable.

Identify the different parameters for the linear regression, and identify the line where the model is applied.

Exercise 2.1.5

Inspect the code associated with this exercise (`ex2_1_5`) and complete the missing lines.

Use the script to evaluate the training and test errors on independent sets. The sets are generated by the same true weight vector and the same noise variance, for two models.

- Identify the number of parameters for the two models
- Discuss the reason why the models perform best for different training set sizes.

Exercise 2.1.6

Compare the training and test errors *per example* as function of the size of the training set for the two models.

Exercise 2.1.7

Compare the values of the training and test errors for large training sets with the value of the noise variance.

2.2 Estimation

This exercise is purely hand–derivation, where we will show three results. It is important to reflect on the results that you derive, don't let yourself be bogged down in pure algebraic manipulations.

Exercise 2.2.1

We will work with the Mean-squared error estimation. Make sure to read sec. 3.9.1 in the book before carrying out this exercise.

We assume the general form of regression

$$\mathrm{y} = g(\mathbf{x}) + \eta$$

where η is zero–mean Gaussian noise with variance σ_η^2. If we have found an optimal estimate in the mean squared sense using data:

$$\hat{g}(\boldsymbol{x}) = \mathbb{E}[\mathrm{y}|\boldsymbol{x}]$$

Then,

$$\mathrm{MSE} = \mathbb{E}\big[(\mathrm{y} - \mathbb{E}[\mathrm{y}|\boldsymbol{x}])^2\big] = \sigma_\eta^2$$

Show the MSE relation. Consider what is the implications of the result?

Exercise 2.2.2 (This is a difficult exercise)

As a first step towards understanding the parameter estimation task, we will analyze the variance of an unbiased estimator. This analysis does not impose any constraints on the model as such, only on the estimator.

From sec 3.5.1 we have the following result that we want to derive.

Let $\hat{\boldsymbol{\theta}}_i, i = 1, 2, \ldots, m$, be unbiased estimators of a parameter vector $\boldsymbol{\theta}$, so that $\mathbb{E}\Big[\hat{\boldsymbol{\theta}}_i\Big] = \boldsymbol{\theta}, i = 1, \ldots, m$. Moreover, assume that the respective estimators are uncorrelated to each other and that all have the same (total) variance, $\sigma^2 = \mathbb{E}\Big[\Big(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}\Big)^T\Big(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}\Big)\Big]$. Then by averaging the estimates,

$$\hat{\boldsymbol{\theta}} = \frac{1}{m}\sum_{i=1}^{m}\hat{\boldsymbol{\theta}}_i$$

the new estimator has total variance:

$$\sigma_c^2 = \mathbb{E}\Big[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\Big] = \frac{1}{m}\sigma^2$$

Show that, by averaging the estimators $\hat{\boldsymbol{\theta}}_i$, we obtain $\sigma_c^2 = \frac{1}{m}\sigma^2$.

Exercise 2.2.3 (This is a difficult exercise)

In this exercise we will provide theoretic proof that a biased estimator can obtain a lower MSE compared to an unbiased estimator. To simplify the derivations, we assume that the estimator is a real scalar instead of a vector.

We define the biased estimator as a scaled estimate of the unbiased estimator,

$$\hat{\theta}_b = (1 + \alpha)\hat{\theta}_u$$

where $\hat{\theta}_u$ is the unbiased estimator. We will show that $\hat{\theta}_b$ has a lower MSE than the corresponding unbiased estimator for certain choices of α, ie. $\mathrm{MSE}(\hat{\theta}_b) < \mathrm{MSE}(\hat{\theta}_u)$.

In addition, we also show that the norm of the biased estimator is lower than the corresponding unbiased estimator, ie. $|\hat{\theta}_b| < |\hat{\theta}_u|$.

The task is the following:

1. Compute $\mathrm{MSE}(\hat{\theta}_b)$ as a function of α, $\mathrm{MSE}(\hat{\theta}_u)$ and θ_0, where θ_0 are the true optimal parameter (ignore the value for α in this step).
2. Show that there exists an α such that $\mathrm{MSE}(\hat{\theta}_b) \leq \mathrm{MSE}(\hat{\theta}_u)$ by placing bounds on α using the expression derived in step 1 . What can you for instance say about the sign of α?
3. Conclude that $\mathrm{MSE}(\hat{\theta}_b) \leq \mathrm{MSE}(\hat{\theta}_u)$ implies that $|1 + \alpha| < 1$ and relate that to $|\hat{\theta}_b|$.

2.3 Bias-variance trade-off

Now we will consider a regularized version of the linear model (Ridge regression), where the parameters $\boldsymbol{\theta}$ are learned by minimization for a fixed value of λ:

$$L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 + \lambda \|\boldsymbol{\theta}\|^2$$

where it is assumed that the bias θ_0 is encoded as part of the $\boldsymbol{x}_n$ vector as in exercise 2.1.2.

Exercise 2.3.1

Use the results from exercise 2.1 to write up the solution for Ridge regression on the form $A\boldsymbol{x} = \boldsymbol{b}$.

Exercise 2.3.2

Inspect the code associated with this exercise (`ex2_3_2`) and complete the requested lines.

Consider why the function is not simultaneously minimized for $(\boldsymbol{\theta}, \lambda)$.

The training set averages generalization error in the point $\mathbf{x}$ can be written as

$$\mathbb{E}_{\mathcal{D}}\left[(f(\mathbf{x}; \mathcal{D}) - \mathbb{E}[\mathrm{y}|\mathbf{x}])^2\right] = \mathbb{E}_{\mathcal{D}}\left[(f(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[f(\mathbf{x}; \mathcal{D})])^2\right] + (\mathbb{E}_{\mathcal{D}}[f(\mathbf{x}; \mathcal{D})] - \mathbb{E}[\mathrm{y}|\mathbf{x}])^2$$

where $\mathbb{E}_{\mathcal{D}}[\cdot]$ is the expectation with respect to training sets.

Hence the average error is split into a variance part, quantifying the variation among solutions for different training sets and a bias part quantifying the performance of the average model with respect to best possible model.

The corresponding code snippet plots the relative amount of variance and bias for a linear model.

Exercise 2.3.3

Plot the average generalization error, the bias error, and the variance error for a large range of weight decay values. Comment on the two regimes where the generalization error stems from variance and bias respectively.

Exercise 2.3.4

Consider the following questions;

- What is the role of the weight decay in these two regimes?
- Which weight decay value would you recommend?

Exercise 2.3.5 (Optional)

In this exercise we will analyze bias-variance trade-off on the example of a simple physical experiment which is tossing the ball up in the air. We consider an experiment in which a ball has been tossed into the air (from the height of an averaged human, 168cm) and its position is

measured as a function of time. The data obtained from this hypothetical experiment is stored in 'h' array.

Inspect the code associated with this exercise (`ex2_3_5`), and answer a following questions:

1. From mechanics we know that the height should vary as the square of the time, so instead of a linear fit to the data we should use a quadratic one:

$$h(t) = h_0 + v_0 t + \frac{1}{2} a t^2$$

   Which degree of plotted polynomials would you choose? Explain why.

2. How does the bias and variance change with the degree of polynomials?

3. In which case do we see underfitting and in which overfitting. Explain why.

4. How does the model complexity change with the degree of polynomials?

HINTS

For exercise 2.2.1

Relate the result to your observations from 2.1.7

For exercise 2.2.2

To solve this problem, first carry out substitution of $\hat{\theta}$ in the expression $\sigma_c^2 = \mathbb{E}[(\hat{\theta}-\theta_o)^T(\hat{\theta}-\theta_o)]$, and then use the result that the estimators are uncorrelated, formally $\mathbb{E}[(\hat{\theta}_i - \theta_o)^T(\hat{\theta}_j - \theta_o)] = \sigma^2\delta_{ij}$, where $\delta_{ij} = 1$ when $i{=}j$ and zero otherwise.

For exercise 2.2.3

To show this, we first need an expression for the MSE of the biased estimator

$$\mathrm{MSE}(\hat{\theta}_b) = \mathbb{E}[(\hat{\theta}_b - \theta_o)^2]$$

Insert by substitution the expression for $\hat{\theta}_b$, and then add $\alpha\theta_o - \alpha\theta_o$ to obtain

$$\mathrm{MSE}(\hat{\theta}_b) = (1+\alpha)^2\mathrm{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2$$

Next we seek the solution for α so that

$$\mathrm{MSE}(\hat{\theta}_b) < \mathrm{MSE}(\hat{\theta}_u)$$

We solve this inequality for α first by substituting the expression for $\mathrm{MSE}(\hat{\theta}_b)$ and then divide both sides with $\theta_o^2 + \mathrm{MSE}(\hat{\theta}_u)$ and rearranging. Note you can assume $\alpha \neq 0$, since $\hat{\theta}_b = \hat{\theta}_u$ for $\alpha = 0$. This will lead to the solution

$$-\frac{2\mathrm{MSE}(\hat{\theta}_u)}{\mathrm{MSE}(\hat{\theta}_u) + \theta_o^2} < \alpha < 0$$

The final step is to verify that the norm of $\hat{\theta}_b < \hat{\theta}_u$. This result is obtained by bounding the lower bound to get the result $|1+\alpha| < 1$.

For exercise 2.3.1

The solution is written using sums in eq. (3.42) in the book.