Drawing Lines

Raghav Saboo

June 2, 2016

Oh the Places You'll Go

Much of machine learning can be boiled down to a function estimation task, and in order to achieve this there are two types of methods adopted: parametric and non-parametric. Parametric methods are where the functional dependence is defined through a fixed number of parameters. On the other hand, non-parametric methods are where the number of parameters that define the function vary with the size of the data set. In both cases the values of the parameters are unknown. These set of notes we will only focus on parametric methods.

Whilst considering parametric methods, we will come across two ways of estimating the unknown values of the parameters. The first set of methods will assume that the unknown parameters have specific "true" values which can be obtained through deterministic steps. The second will take a probabilistic approach, wherein the unknown parameters will be treated as random variables. Thus probability distributions will be used to describe the input and output variables, and thus removing the need to find specific values for the unknown parameters.

As we take a detailed tour of the different approaches to parameter estimation we will look singularly at linear functions. The fact that this is not limiting our ability to generalize to non-linear models will become apparent as we progress further through the material. Ultimately the idea is to create a clear map of the vast landscape of parameter estimation which can otherwise be so hard to comprehend as a student.

Contents

Chapter 1

Basic Concepts

1.1 General Setup

- We will start by *adopting* a form, such as a linear quadtratic function, with the goal of estimating the associated unknown coefficients so that the function matches the spacial arrangement of the data as closely as possible.
- Given a set of data points $(y_n, \boldsymbol{x}_n)$, $y_n \in \mathbb{R}$, $\boldsymbol{x}_n \in \mathbb{R}^l$, $n = 1, 2, ..., N$, and a parametric set of functions,

$$\mathcal{F} := \{f_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} \in \mathcal{A} \subseteq \mathbb{R}^K\} \tag{1.1}$$

  find a function in $\mathcal{F}$, which will be denoted as $f(\cdot) := f_{\boldsymbol{\theta}_*}(\cdot)$, such that given a value of $\boldsymbol{x} \in \mathbb{R}^l$, $f(\boldsymbol{x})$ best approximates the corresponding value $y \in \mathbb{R}$. The value $\boldsymbol{\theta}_*$ is the value that results from the estimation procedure.
- The choice of $\mathcal{F}$ has to be based on as much a priori information as possible concerning the physical mechanism that underlies the generation of the data. The most common approach is to iterate over different families of functions and evaluate the performance of each according to a chosen criterion.
- Having adopted a parametric family of functions, $\mathcal{F}$, one has to get an estimate for the unknown set of parameters. A *non-negative loss* function is usually adopted, which quantifies the deviation/error between

the measured value of y and the predicted one using the corresponding measurements $\boldsymbol{x}$, as in $f_{\boldsymbol{\theta}}(x)$.

$$\mathcal{L}(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \longmapsto [0,\infty), \tag{1.2}$$

and compute $\boldsymbol{\theta}_*$ so as to minimize the total loss,

$$f(\cdot) := f_{\boldsymbol{\theta}_*}(\cdot) : \boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}\in\mathcal{A}} J(\boldsymbol{\theta}) \tag{1.3}$$

assuming that a minimum exists.

- There may be more than one optimal values $\boldsymbol{\theta_*}$, depending on the shape of $J(\boldsymbol{\theta})$.
- A common combination used to introduce estimation techniques is the linear class of functions with a least squares (LS) loss function,

$$\mathcal{L}(y, f_\theta(\boldsymbol{x})) = (y - f_\theta(\boldsymbol{x}))^2, \tag{1.4}$$

and there are good very good reasons for this.

1. Linearity with the LS loss function turns out to simplify the algebra and hence allows one to understand the various "secrets" that underlie the area of parameter estimation.
2. Understanding linearity is very important. Treating nonlinear tasks, most often, turns out to finally resort to a linear problem. A good example of this is the transformation

$$\mathbb{R} \ni x \longmapsto \boldsymbol{\phi}(x) := \begin{bmatrix} x \\ x^2 \end{bmatrix} \in \mathbb{R} \tag{1.5}$$

which is the same as

$$y = \theta_0 + \theta_1\phi_1(x) + \theta_2\phi_2(x) \tag{1.6}$$

which is a linear model with respect to the components $\phi_k(x), k = 1, 2$, of the two-dimension image, $\boldsymbol{\phi}(x)$, of x. This simple trick is at the heart of a number of nonlinear methods that will be treated later on in these notes.

1.2 Introduction to Linear Regression

- Regression involves the task of modeling the relationship of a dependent random variable, y, which is considered to be the response of a system to changes in independent variables, $x_1, x_2, ..., x_l$.
- The independent variables will be represented as the components of an equivalent random vector $\boldsymbol{x}$.
- The relationship is modeled via an additive disturbance or noise term, η.

$$y = \theta_0 + \theta_1 x_1 + ... + \theta_l x_l + \eta = \boldsymbol{\theta}^T \boldsymbol{x} + \eta \tag{1.7}$$

- The noise variable, η, is an unobserved random variable, and the goal of the regression task is to estimate the parameter vector, $\boldsymbol{\theta}$, given a set of data, $(y_n, \boldsymbol{x}_n)$ where $n = 1, 2, ..., N$. This is also known as the training data set.
- We can state that the prediction model for a $\hat{\boldsymbol{\theta}}$ determined using training data set, is:

$$\hat{y} = \hat{\boldsymbol{\theta}}^T x \tag{1.8}$$

  where $\hat{\boldsymbol{\theta}}$ can be set to $\boldsymbol{\theta}_*$ obtained from minimizing the LS loss function

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 \tag{1.9}$$

- Taking the derivative with respect to $\boldsymbol{\theta}$ and equating to the zero vector $\mathbf{0}$ results in

$$\left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T \right) \hat{\boldsymbol{\theta}} = \sum_{n=1}^{N} \boldsymbol{x}_n y_n \tag{1.10}$$

  which is equivalent to

$$X^T X \hat{\boldsymbol{\theta}} = X^T \boldsymbol{y} \tag{1.11}$$

  and the LS estimate is

$$\hat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \boldsymbol{y} \tag{1.12}$$

  assuming that $(X^T X)^{-1}$ exists.

This solution is unique, provided that the X^TX is invertible. The uniqueness is due to the strictly convex parabolic shape the LS loss function.

1.3 Biased versus Unbiased Estimation

- In supervised learning, we are given a set of training points, $(y_n, \boldsymbol{x}_n)$, $n = 1, 2, ..., N$, and we return an estimate of the unknown paramter vector, say $\hat{\boldsymbol{\theta}}$.
- However, the training points are random variables (in the deterministic world), and thus if we are given another set of N observations of the same random variables, these are going to be different, and obviously the resulting estimate will also be different. In other words, by changing our training data different estimates result.
- An estimate, such as $\hat{\boldsymbol{\theta}}$, has a specific value, which is the result of a function acting on a set of observations, on which our chosen estimate depends. In general, we could generalize and write that

$$\hat{\boldsymbol{\theta}} = f(\boldsymbol{y}, X). \tag{1.13}$$

- Once we allow the set of observations to change randomly, and the estimate becomes itself a random variable, we write the previous equation in terms of the corresponding random variables,

$$\hat{\Theta} = f(\boldsymbol{y}, X), \tag{1.14}$$

  and we refer to this functional dependence as the estimator of the unknown vector $\boldsymbol{\theta}$.
- In order to simplify the analysis and focus on the insight behind the methods, we will assume that our parameter space is that of real numbers, $\mathbb{R}$. We will also assume that the model (i.e., the set of functions $\mathcal{F}$), which we have adopted for modeling our data, is the correct one and the value of the associated true parameter is equal to θ_0 (unknown to us).

- Let $\hat{\Theta}$ denote the random variable of the associated estimator; then the squared error loss function to quantify deviations, a reasonable criterion to measure the performance of an estimate is the *mean-square error* (MSE),

$$MSE = \mathbb{E}\left[\left(\hat{\theta} - \theta_0\right)^2\right] \tag{1.15}$$

  where the mean $\mathbb{E}$ is taken over all possible training data sets of size N. If the MSE is small, then we expect that, on average, the resulting estimates will be close to the true value. Additionally, if we insert the mean value $\mathbb{E}[\hat{\theta}]$ of $\hat{\theta}$ into (1.15) to get

$$\begin{aligned} MSE &= \mathbb{E}\left[\left(\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right) + \left(\mathbb{E}[\hat{\theta}] - \theta_0\right)\right)^2\right] \\ &= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] + \left(\mathbb{E}[\hat{\theta}] - \theta_0\right)^2 \end{aligned} \tag{1.16}$$

  where first term is $Variance$ around the mean value, and the second term is $Bias^2$: the deviation of the mean value of the estimator from the true one.

- One may think that choosing an estimator that is *unbiased*, as is $\mathbb{E}[\hat{\theta}] = \theta_0$, such that the second term in (1.16) becomes zero, is a reasonable choice. Assume that we have L different training sets, each comprising N points. Let us denote each data set by $\mathcal{D}_i, i = 1, 2, ..., L$. For each one, an estimate $\hat{\theta}_i$, $i = 1, 2, ..., L$ will result. Then, form the new estimator by taking the average value,

$$\hat{\theta}^{(L)} := \frac{1}{L}\sum_{i=1}^{L}\hat{\theta}_i \tag{1.17}$$

  This is also an unbiased estimator, because

$$\mathbb{E}[\hat{\theta}^{(L)}] = \frac{1}{L}\sum_{i=1}^{L}\mathbb{E}[\hat{\theta}_i] = \theta_0. \tag{1.18}$$

- Moreover, assuming that the involved estimators are mutually uncorrelated,

$$\mathbb{E}\left[(\hat{\theta}_i - \theta_0)(\hat{\theta}_j - \theta_0)\right] = 0, \tag{1.19}$$

and of the same variance, σ^2, then the variance of the new estimator is now much smaller.

$$\sigma^2_{\hat{\theta}^{(L)}} = \mathbb{E}\left[\left(\hat{\theta}^{(L)} - \theta_0\right)^2\right] = \frac{\sigma^2}{L} \tag{1.20}$$

- Hence, by averaging a large number of such unbiased estimators, we expect to get an estimate close to the true value. However, in practice, data is not always abundant. As a matter of fact, very often the opposite is true and one has to be very careful about how to exploit it. In such cases, where one cannot afford to obtain and average a large number of estimators, an unbiased estimator may not necessarily be the best choice.

- Also going back to (1.16), there is no reason to suggest that by making the second term equal to zero, the MSE becomes minimum.

- Instead of computing the MSE for a given estimator, let us replace $\hat{\theta}$ with θ in (1.16) and compute an estimator that will minimize the MSE with respect to θ, directly.

- In this case, focusing on unbiased estimators, $\mathbb{E}[\theta] = \theta_0$, introduces a constraint to the task of minimizing the MSE, and it is well-known that an unconstrained minimization problem always results in loss function values that are less than or equal to any value generated by a constrained counterpart,

$$\min_{\theta} MSE(\theta) \leq \min_{\theta:\mathbb{E}[\theta]=\theta_0} MSE(\theta) \tag{1.21}$$

where the dependence of MSE on the estimator θ in (1.21) is explicitly denoted.

- Let us denote by $\hat{\theta}_{MVU}$ a solution of the task $\min_{\theta:\mathbb{E}[\theta]=\theta_0} MSE(\theta)$, i.e. the unbiased estimator.

- Motivated by (1.21), our next goal is to search for a biased estimator, which results, hopefully, in a smaller MSE. Let us denote this estimator as $\hat{\theta}_b$.

- For the sake of illustration, and in order to order to limit our search for $\hat{\theta}_b$, we consider here only $\hat{\theta}_b$s that are scalar multiples of $\hat{\theta}_{MVU}$, so that

$$\hat{\theta}_b = (1+\alpha)\hat{\theta}_{MVU}, \tag{1.22}$$

  where $\alpha \in \mathbb{R}$ is a free parameter. Notice that $\mathbb{E}[\hat{\theta}_b] = (1+\alpha)\theta_0$. By substituting (1.22) into (1.15) and after some simple algebra we obtain

$$MSE(\hat{\theta}_b) = (1+\alpha)^2 MSE(\hat{\theta}_{MVU}) + \alpha^2\theta_0^2. \tag{1.23}$$

- In order to get $MSE(\hat{\theta}_b) < MSE(\hat{\theta}_{MVU}), \alpha$ must be in the range

$$-\frac{MSE(\hat{\theta}_{MVU})}{MSE(\hat{\theta}_{MVU}) + \theta_\sigma^2} < \alpha < 0. \tag{1.24}$$

- It is easy to verify that the previous range implies that $|1+\alpha| < 1$. Hence, $|\hat{\theta}_b| = |(1+\alpha)\hat{\theta}_{MVU}| < |\hat{\theta}_{MVU}$. We can go a step further and try to compute the optimum value of α, which corresponds to the minimum MSE. By taking the derivative of $MSE(\hat{\theta}_b)$ in (1.23) with respect to α, it turns out that this occurs for

$$\alpha_* = -\frac{MSE(\hat{\theta}_{MVU})}{MSE(\hat{\theta}_{MVU}) + \theta_0^2} = -\frac{1}{1 + \frac{\theta_0^2}{MSE(\hat{\theta}_{MVU})}} \tag{1.25}$$

- Therefore, we have found a way to obtain the optimum estimator, among those in the set $\hat{\theta}_b = (1+\alpha)\hat{\theta}_{MVU} : \alpha \in \mathbb{R}$, which results in minimum MSE. This is true, but it is not realizable as the optimal value of α is given in terms of the unknown θ_0. However it does show one important result: If we want to do better than the MVU, then, one way is to **shrink** the norm of the MVU estimator.

- Shrinking the norm is a way of introducing bias into an estimator.

- What we have said so far is readily generalized to parameter vectors. An unbiased parameter vector satisfies

$$\mathbb{E}[\Theta] = \boldsymbol{\theta}_0,$$

and the MSE around the true value, $\boldsymbol{\theta}_0$, is defined as

$$MSE = \mathbb{E}\left[(\Theta - \boldsymbol{\theta}_0)^T(\Theta - \boldsymbol{\theta}_0)\right]$$

- Looking carefully at the previous definition reveals that the MSE for a paramter vector is the sum of the MSEs of the components, $\theta_i, i = 1, 2, ..., l$, around the corresponding true values θ_{0i}.

Cramér-Rao Lower Bound

- The Cramér-Rao lower bound is an elegant theorem and one of the most well-known techniques used in statistics. It provides a lower bound on the variance of any unbiased estimator.
- This is very important because:
  1. It offers the means to assert whether an unbiased estimator has minimum variance, which, of course, in this case coincides with the corresponding MSE in (1.15).
  2. And if the above is not the case, it can be used to indicate how far away the performance of an unbiased estimator is from the optimal one.
  3. Finally it provides the designer with a tool to known the best possible performance that can be achieved by an unbiased estimator.
- We are looking for a bound of the variance of an unbiased estimator, whose randomness is due to the randomness of the training data.

  Theorem 1 (Cramér-Rao Bound)**.** *Let $\boldsymbol{x}$ denote a random vector and let $\mathcal{X} = \boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}$ denote the set of N observations, corresponding to a random vector. The corresponding joint pdf is parameterized in terms of the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^l$. The log-likelihood is then defined as,*

$$L(\boldsymbol{\theta}) := \log p(\mathcal{X}; \boldsymbol{\theta}).$$

Define the Fisher's Information Matrix as

$$J = \begin{bmatrix} \mathbb{E}\left[\frac{\partial^2 \log p}{\partial \theta_1^2}\right] & \mathbb{E}\left[\frac{\partial^2 \log p}{\partial \theta_1 \partial \theta_2}\right] & \cdots & \mathbb{E}\left[\frac{\partial^2 \log p}{\partial \theta_1 \partial \theta_l}\right] \\ \vdots & \vdots & \cdots & \vdots \\ \mathbb{E}\left[\frac{\partial^2 \log p}{\partial \theta_l \partial \theta_1}\right] & \mathbb{E}\left[\frac{\partial^2 \log p}{\partial \theta_l \partial \theta_2}\right] & \cdots & \mathbb{E}\left[\frac{\partial^2 \log p}{\partial \theta_l^2}\right] \end{bmatrix} \tag{1.26}$$

Let $I := J^{-1}$ and let $I(i,i)$ denote the ith diagonal element of I. If $\hat{\theta}_i$ is any unbiased estimator of the ith component, θ_i, of $\boldsymbol{\theta}$, then the corresponding variance of the estimator,

$$\sigma^2_{\hat{\theta}_i} \geq I(i,i). \tag{1.27}$$

This is known as the Cramér-Rao lower bound, and if an estimator achieves this bound it is said to be efficient and it is unique.

- Moreover, the necessary and sufficient condition for obtaining an unbiased estimator that achieves the bound is the existence of a function $g(\cdot)$ such that for all possible values of θ_i,

$$\frac{\partial \log p(\mathcal{X};\theta)}{\partial \theta_i} = I(\theta)(g(\mathcal{X} - \theta)). \tag{1.28}$$

The MVU estimate is then given by

$$\hat{\theta}_i = g(\mathcal{X}) := g(\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}) \tag{1.29}$$

and the variance of the respective estimator is equal to $\frac{1}{I(\theta_i)}$

Applied to Linear Regression

- Lets consider a simple linear regression model as seen below

$$y_n = \theta x + \eta_n$$

- To simplify the discussion, we assume that our N observations are the result of different realizations of the noise variable ONLY (i.e. fixed input x).

- Further assume that η_n are samples of a Gaussian white noise with zero mean and variance equal to σ_η^2
- The joint pdf of the output observations is given by

$$p(y;\theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp -\frac{(y_n-\theta)^2}{2\sigma_\eta^2}$$

- We can derive the corresponding Cramér-Rao bound as follows:

$$\frac{\partial \log p(y;\theta)}{\partial\theta} = \frac{1}{\sigma_\eta^2}\sigma_{n=1}N(y_n-\theta) = \frac{N}{\sigma_\eta^2}(\bar{y}-\theta)$$

  where

$$\bar{y} := \frac{1}{N}\sum_{n=1}^{N} y_n$$

  and the second derivative as required by the Cramér-Rao theorem

$$\frac{\partial^2 \log p(y;\theta)}{\partial\theta^2} = -\frac{N}{\sigma_\eta^2}.$$

- Hence

$$I(\theta) = \frac{N}{\sigma_\eta^2}$$

  and an efficient estimator would be

$$\sigma_{\hat{\theta}}^2 \geq \frac{\sigma_\eta^2}{N}$$

- We can easily verify that the corresponding estimator, $\bar{y}$ is indeed an unbiased one

$$\mathbb{E}[\bar{y}] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[y_n] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\theta+\eta_n] = \theta$$

- For this particular task and having assumed that the noise is Gaussian, the LS estimator is equal to the MVU estimator and it attains the Cramér-Rao bound.

- However, if the input is not fixed, but also varies from experiment to experiment and the training data become (y_n, x_n), then the LS estimator attains the Cramér-Rao bound only asymptotically, for large values of N.
- Also if the assumptions for the noise being Gaussian **and** white are not valid, then the LS estimator is not efficient anymore.

Sufficient Statistic

- If an efficient estimator does not exist, this does not necessarily mean that the MVU estimator cannot be determined.
- The notion of **sufficient statistic** and the Rao-Blackwell theorem help us here. Given a random vector, $\mathbf{x}$ which depends on parameter θ, sufficient statistic for the unknown parameter is a function

  $$T(\mathcal{X}) := T(\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N})$$

  of the respective observations, which contains **all** information about θ.
- From a mathematical point of view, a statistic $T(\mathcal{X}$ is said to be sufficient for the paramter θ if the conditional joint pdf

  $$p(\mathcal{X}|T(\mathcal{X}; \theta),$$

  does not depend on θ and thus $T(\mathcal{X})$ must provide *all* information about θ which is contained in the set $\mathcal{X}$.
- Once $T(\mathcal{X})$ is known, $\mathcal{X}$ is no longer needed: hence the name "sufficient statistic" i.e. no more information can be extracted from $\mathcal{X}$.
- In the case of parameter vectors $\boldsymbol{\theta}$, the sufficient statistic may be a *set* of functions, called a *jointly sufficient statistic.*

  Theorem 2 (Factorization Theorem)**.** *A statistic $T(\mathcal{X})$ is sufficient if and only if the respective joint pdf can be factorized as*

  $$p(\mathcal{X}; \boldsymbol{\theta}) = h(\mathcal{X})g(T(\mathcal{X}), \boldsymbol{\theta}).$$

  One part of the factorization only depends on the statistic and parameters, and a second part that is independent of the parameters.

Applied to Linear Regression

- Let x be a Gaussian, $\mathcal{N}(\mu, \sigma^2)$, random variable and let the set of observations be $\mathcal{X} = x_1, x_2, ..., x_N$. Assume μ to be the unknown parameter.
- The joint pdf is given by

$$p(\mathcal{X}; \mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp -\frac{\sum_{n=1}^{N}(x_n - \mu)^2}{2\sigma^2}$$

- Given the identity

$$\sum_{n=1}^{N}(x_n - \mu)^2 = \sum_{n=1}^{N}(x_n - S_\mu)^2 + N(S_\mu - \mu)^2$$

  the joint pdf becomes

$$p(\mathcal{X}; \mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp -\frac{\sum_{n=1}^{N}(x_n - S_\mu)^2}{2\sigma^2} \exp -\frac{\sum_{n=1}^{N} N(S_\mu - \mu)^2}{2\sigma^2}$$

- Similarly if the unknown parameter is the variance σ^2 then the sufficient statistic is

$$\bar{S_\sigma^2} := \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

- If both μ and σ^2 are unkown then the sufficient statistic is the set (S_μ, S_σ^2) where

$$S_\sigma^2 := \frac{1}{N}\sum_{n=1}^{N}(x_n - S_\mu)^2$$

1.4 Regularization

- One approach to improve the performance of an estimator is to shrink the norm of the MVU estimator. *Regularization* is a mathematical tool to impose a priori information on the structure of the solution, which comes as the outcome of an optimization task.

- We can reformulate the LS minimization task as

$$\text{minimize:} \quad J(\boldsymbol{\theta}) = \sum_{n=1}^{N}(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 \tag{1.30}$$

$$\text{subject to:} \quad ||\boldsymbol{\theta}||^2 \leq \rho \tag{1.31}$$

  where $||\cdot||$ is the Euclidean norm of a vector.

- Here we do not allow the LS criterion to be completely "free" to reach a solution, but we limit the space in which to search for it. The optimal value of ρ cannot be analytically obtained and thus we have to experiment in order to select an estimator that has good performance.

- Thus the optimization task for a LS loss function can be written as

$$\text{minimize:} \quad L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N}(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n) + \lambda||\boldsymbol{\theta}||^2 \tag{1.32}$$

  which is often referred to as Ridge Regression.

- The specific choices of $\lambda \geq 0$ and ρ are equivalent tasks.

- Taking the gradient of L in the equation above with respect to $\boldsymbol{\theta}$ results in the following solution:

$$\left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T + \lambda I\right) \hat{\boldsymbol{\theta}} = \sum_{n=1}^{N} y_n x_n \tag{1.33}$$

- The presence of λ biases the new solution away from that which would have been obtained from the unregularized LS formulation. Thus Ridge Regression aims to reduce the norm of the estimated vector *at the same time* as trying to keep the sum of squared errors small.

- This is achieved by modifying the vector components, θ_i, so as to reduce the contribution in the misfit measuring term from less informative directions in the input space.

- That is reducing the norm can be considered as an attempt to "simplify" the structure of the estimator, because smaller number of components of the regressor now have an important say.

- Other regularizers can be used in place of the Euclidean norm, such as the ℓ_p norms with $p \geq 1$.
- In practice the bias paramter, θ_0 is left out from the norm in the regularization term; penalization of the intercept would make the procedure dependent on the origin chosen for y.
- Reducing the norm can be considered as an attempt to "simplify" the structure of an estimator, because a smaller number of components of the regressor now have an important say. This viewpoint becomes more clear if one considers nonlinear models.

Inverse problems: Ill-conditioning and overfitting

- Most tasks in machine learning belong to the so-called *inverse problems*. This encompasses all the problems where one has to infer/predict/estimate the values of a model based on a set of available input/output observations-training data.
- In a less mathematical terminology, inverse problems unravel unkown causes from known effects; in other words, to reverse the cause-effect relations.
- Inverse problems are typically ill-posed, as opposed to the well-posed ones. Well-posed problems are characterized by (a) the existence of a solution, (b) the uniqueness of the solutions and (c) the stability of the solution.
- In machine learning problems, the obtained solution may be very sensitive to changes of the training set. *Ill conditioning* is another term used to describe this sensitivity.
- The reson for ill-conditioning is that the model used to describe the data can be complex; i.e. the number of the unknown free parameters is large with respect to the number of data points. This is also known as *overfitting*.
- *Overfitting* means that the estimated parameters of the unknown model learn too much about the idiosyncrasies of the specific training

data set, and the model performs badly when it deals with another set of data, other than that used for the training.

- Regularization is an elegant and efficient tool to cope with the complexity of a model; that is, to make it less complex, and more smooth.
- When dealing with more complex, compared to linear, models, one can use constraints on the smoothness of the involved nonlinear function; for example, by involving derivatives of the model function in the regularization term.
- Examples:

  1. In the LS linear regression task, if the number, N, of the training points is less than the dimension of the regressors $\mathbf{x}_n$, then the $\ell \times \ell$ matrix, $\bar{\sum} = \sum_n x_n x_n^T$, is not invertible. Indeed, each term in the summation is the outer product of a vector with itself and hence it is a matrix of rank one. Thus, as we know from linear algebra, we need at least ℓ linearly independent terms of such matrices to guarantee that the sum is of full rank, hence invertible. In ridge regression however, this can be bypassed, because of the presence of λI guarantees that the matrix in invertible.
  2. Another example where regularization can help to obtain a solution, and, more important, a unique solution to an otherwise unsolvable problem, is when the model's order is large compared to the number of data, albeit we know that it is sparse. That is, only a very small percentage of the model's paramters are nonzero - here LS linear regression approach has no solution. However, regularizing the LS loss function using the ℓ_1 norm of the parameters' vector can lead to a unique solution; the ℓ_1 norm of a vector comprises the sum of absolute values of its components.

- Regularization is also closely related tot he task of using priors in Bayesian learning.

Chapter 2

Deterministic Methods

2.1 Mean Squared Error

The Normal Equations

The Loss Function Surface

MSE Linear Estimator

Geometric Viewpoint: Orthogonality Condition

Gauss Markov Theorem

Stochastic Gradient Descent

2.2 Least Squares

Geometric Perspective

Statistical Properties

Orthogonalizing the Column Space of X: the SVD Method

Recursive Least Squares

Comparison with Stochastic Gradient Descent

2.3 Convex Analysis

Applied to Linear Regression

2.4 Reproducing Kernel Hilbert Spaces

Chapter 3

Bayesian (Probabilistic) Methods

3.1 Exponential Family of Probability Distributions

The Maximum Entropy

3.2 Maximum Likelihood

3.3 Maximum a Posteriori (MAP) Probability Estimation

3.4 The Curse of Dimensionality

3.5 The EM Algorithm

The Evidence Function

Laplacian Approximation of the Evidence Function

Latent Variables and the EM Algorithm

Lower Bound Maximization View

Linear Regression using EM Algorithm

Mixing Linear Regression Models