Chapter 7
Linear Regression

7.1 Introduction

Linear regression is the work horse of statistics and (supervised) machine learning. When augmented with kernels or other forms of basis function expansion, it can model also nonlinear relationships. And when the Gaussian output is replaced with a Bernoulli or multinoulli distribution, it can be used for classification, as we will see below. So it pays to study this model in detail.

7.2 Representation

$$p(y|\boldsymbol{x},\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x},\sigma^2) \tag{7.1}$$

where $\boldsymbol{w}$ and $\boldsymbol{x}$ are extended vectors, $\boldsymbol{x} = (1,x)$, $\boldsymbol{w} = (b,w)$.

Linear regression can be made to model non-linear relationships by replacing $\boldsymbol{x}$ with some non-linear function of the inputs, $\phi(\boldsymbol{x})$

$$p(y|\boldsymbol{x},\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^T\phi(\boldsymbol{x}),\sigma^2) \tag{7.2}$$

This is known as **basis function expansion**. (Note that the model is still linear in the parameters $\boldsymbol{w}$, so it is still called linear regression; the importance of this will become clear below.) A simple example are polynomial basis functions, where the model has the form

$$\phi(x) = (1,x,\cdots,x^d) \tag{7.3}$$

7.3 MLE

Instead of maximizing the log-likelihood, we can equivalently minimize the **negative log likelihood** or **NLL**:

$$\text{NLL}(\boldsymbol{\theta}) \triangleq -\ell(\boldsymbol{\theta}) = -\log(\mathcal{D}|\boldsymbol{\theta}) \tag{7.4}$$

The NLL formulation is sometimes more convenient, since many optimization software packages are designed to find the minima of functions, rather than maxima.

Now let us apply the method of MLE to the linear regression setting. Inserting the definition of the Gaussian into the above, we find that the log likelihood is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2\right)\right] \tag{7.5}$$

$$= -\frac{1}{2\sigma^2}\text{RSS}(\boldsymbol{w}) - \frac{N}{2}\log(2\pi\sigma^2) \tag{7.6}$$

RSS stands for **residual sum of squares** and is defined by

$$\text{RSS}(\boldsymbol{w}) \triangleq \sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 \tag{7.7}$$

We see that the MLE for $\boldsymbol{w}$ is the one that minimizes the RSS, so this method is known as **least squares**.

Let's drop constants wrt $\boldsymbol{w}$ and NLL can be written as

$$\text{NLL}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 \tag{7.8}$$

There two ways to minimize $\text{NLL}(\boldsymbol{w})$.

7.3.1 OLS

Define $\boldsymbol{y} = (y_1, y_2, \cdots, y_N)$, $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}$, then $\text{NLL}(\boldsymbol{w})$ can be written as

$$\text{NLL}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{7.9}$$

When $\mathcal{D}$ is small(for example, $N < 1000$), we can use the following equation to compute $\boldsymbol{w}$ directly

$$\hat{\boldsymbol{w}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{7.10}$$

The corresponding solution $\hat{\boldsymbol{w}}_{\text{OLS}}$ to this linear system of equations is called the **ordinary least squares** or **OLS** solution.

Proof. We now state without proof some facts of matrix derivatives (we wont need all of these at this section).

$$trA \triangleq \sum_{i=1}^{n} A_{ii}$$

$$\frac{\partial}{\partial A} AB = B^T \tag{7.11}$$

$$\frac{\partial}{\partial A^T} f(A) = \left[\frac{\partial}{\partial A} f(A)\right]^T \tag{7.12}$$

$$\frac{\partial}{\partial A} ABA^T C = CAB + C^T AB^T \tag{7.13}$$

$$\frac{\partial}{\partial A} |A| = |A|(A^{-1})^T \tag{7.14}$$

Then,

$$\begin{aligned}
\text{NLL}(\boldsymbol{w}) &= \frac{1}{2N}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}) \\
\frac{\partial \text{NLL}}{\boldsymbol{w}} &= \frac{1}{2}\frac{\partial}{\boldsymbol{w}}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{y}^T\boldsymbol{y}) \\
&= \frac{1}{2}\frac{\partial}{\boldsymbol{w}}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w}) \\
&= \frac{1}{2}\frac{\partial}{\boldsymbol{w}} tr(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w}) \\
&= \frac{1}{2}\frac{\partial}{\boldsymbol{w}}(tr\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2tr\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w})
\end{aligned}$$

Combining Equations 7.12 and 7.13, we find that

$$\frac{\partial}{\partial A^T} ABA^T C = B^T A^T C^T + BA^T C$$

Let $A^T = \boldsymbol{w}, B = B^T = \boldsymbol{X}^T\boldsymbol{X}$, and $C = I$, Hence,

$$\begin{aligned}
\frac{\partial \text{NLL}}{\boldsymbol{w}} &= \frac{1}{2}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{X}^T\boldsymbol{y}) \\
&= \frac{1}{2}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}) \\
\frac{\partial \text{NLL}}{\boldsymbol{w}} &= 0 \Rightarrow \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y} = 0 \\
\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} &= \boldsymbol{X}^T\boldsymbol{y} \\
\hat{\boldsymbol{w}}_{\text{OLS}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned} \tag{7.15}$$

Equation 7.15 is known as the **normal equation**.

7.3.1.1 Geometric interpretation

See Figure 7.1.

To minimize the norm of the residual, $\boldsymbol{y} - \hat{\boldsymbol{y}}$, we want the residual vector to be orthogonal to every column of $\boldsymbol{X}$,so $\tilde{\boldsymbol{x}}_j(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0$ for $j = 1 : D$. Hence

$$\begin{aligned}
\tilde{\boldsymbol{x}}_j(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0 &\Rightarrow \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0 \\
&\Rightarrow \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned} \tag{7.16}$$

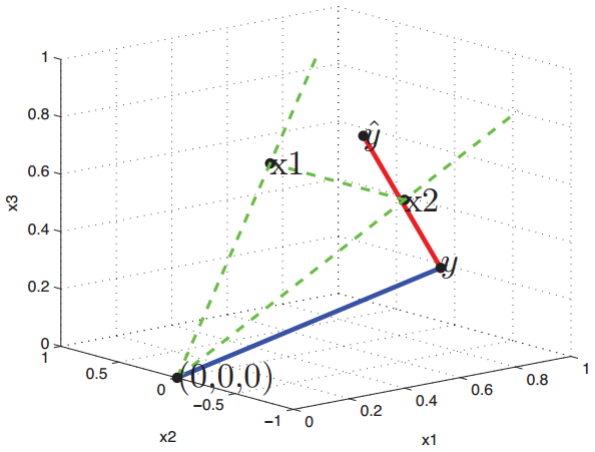

Fig. 7.1: Graphical interpretation of least squares for $N = 3$ examples and $D = 2$ features. $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_2$ are vectors in $\mathbb{R}^3$; together they define a 2D plane. $\boldsymbol{y}$ is also a vector in $\mathbb{R}^3$ but does not lie on this 2D plane. The orthogonal projection of $\boldsymbol{y}$ onto this plane is denoted $\hat{\boldsymbol{y}}$. The red line from $\boldsymbol{y}$ to $\hat{\boldsymbol{y}}$ is the residual, whose norm we want to minimize. For visual clarity, all vectors have been converted to unit norm.

7.3.2 SGD

When $\mathcal{D}$ is large, use stochastic gradient descent(SGD).

$$\because \frac{\partial}{\partial w_i}\text{NLL}(\boldsymbol{w}) = \sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)x_{ij} \tag{7.17}$$

$$\begin{aligned}
\therefore w_j &= w_j - \alpha\frac{\partial}{\partial w_j}\text{NLL}(\boldsymbol{w}) \\
&= w_j - \sum_{i=1}^{N}\alpha(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)x_{ij}
\end{aligned} \tag{7.18}$$

$$\therefore \boldsymbol{w} = \boldsymbol{w} - \alpha(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)\boldsymbol{x} \tag{7.19}$$

7.4 Ridge regression(MAP)

One problem with ML estimation is that it can result in overfitting. In this section, we discuss a way to ameliorate this problem by using MAP estimation with a Gaussian prior.

7.4.1 Basic idea

We can encourage the parameters to be small, thus resulting in a smoother curve, by using a zero-mean Gaussian prior:

$$p(\boldsymbol{w}) = \prod_j \mathcal{N}(w_j|0,\tau^2) \tag{7.20}$$

where $1/\tau^2$ controls the strength of the prior. The corresponding MAP estimation problem becomes

$$\arg\max_{\boldsymbol{w}} \sum_{i=1}^{N} \log \mathcal{N}(y_i|w_0+\boldsymbol{w}^T\boldsymbol{x}_i,\sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j|0,\tau^2) \tag{7.21}$$

It is a simple exercise to show that this is equivalent to minimizing the following

$$J(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}(y_i-(w_0+\boldsymbol{w}^T\boldsymbol{x}_i))^2 + \lambda\|\boldsymbol{w}\|^2, \lambda \triangleq \frac{\sigma^2}{\tau^2} \tag{7.22}$$

Here the first term is the MSE/ NLL as usual, and the second term, $\lambda \geq 0$, is a complexity penalty. The corresponding solution is given by

$$\hat{\boldsymbol{w}}_{\text{ridge}} = (\lambda \boldsymbol{I}_D + \boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{7.23}$$

This technique is known as **ridge regression**,or **penalized least squares**. In general, adding a Gaussian prior to the parameters of a model to encourage them to be small is called ℓ_2 **regularization** or **weight decay**. Note that the offset term w_0 is not regularized, since this just affects the height of the function, not its complexity.

We will consider a variety of different priors in this book. Each of these corresponds to a different form of **regularization**. This technique is very widely used to prevent overfitting.

7.4.2 Numerically stable computation *

$$\hat{\boldsymbol{w}}_{\text{ridge}} = \boldsymbol{V}(\boldsymbol{Z}^T\boldsymbol{Z} + \lambda \boldsymbol{I}_N)^{-1}\boldsymbol{Z}^T\boldsymbol{y} \tag{7.24}$$

7.4.3 Connection with PCA *

7.4.4 Regularization effects of big data

Regularization is the most common way to avoid overfitting. However, another effective approach which is not always available is to use lots of data. It should be intuitively obvious that the more training data we have, the better we will be able to learn.

In domains with lots of data, simple methods can work surprisingly well (Halevy et al. 2009). However, there are still reasons to study more sophisticated learning methods, because there will always be problems for which we have little data. For example, even in such a data-rich domain as web search, as soon as we want to start personalizing the results, the amount of data available for any given user starts to look small again (relative to the complexity of the problem).

7.5 Bayesian linear regression

TODO