A Summary of Machine Learning Algorithms in Elements of Statistical Learning

Dr. Lei Fang

University of St Andrews

Abstract. This article summarises the techniques discussed in the seminal book Elements of Statistical Learning (ESL) [1]. A wide range of algorithms and techniques are presented in the book roughly following the order of each chapter's topic. Therefore, a summary of all the techniques provides another way to summarize the book. The summary also provides a quick reference for readers to compare and choose the appropriate techniques for their application.

Table of Contents

1 Introduction

The ESL book, consisting of 18 chapters in total, has covered a very wide range of statistical learning techniques at various depth. The presentation of these techniques are scattered in the 18 chapters based on the main theme of each chapter. For example, Chapter 4 is devoted to linear methods for classification, Chapter 6 focuses on kernel smoothing methods and its application in density estimation, while Chapter 8 and 14 talks mainly about general model inference algorithm and unsupervised learning algorithms. And some techniques are repetitively mentioned at various places in the book. Take mixture model as an example: it is first introduced in Chapter 4 as a classifier, namely Linear/Quadratic Discriminant Analysis (LDA, QDA), then in Chapter Six for its application in density estimation, and Chapter 8 as a vehicle model to introduce EM algorithm and Gibbs sampling, and Chapter 14 as a clustering model (K-means). These techniques, although widely known by their own names for various different applications, actually share the same statistical model behind the scene. This article aims to link those techniques mentioned in the book together by tracing back to their original model. We believe this is the best way to compare and contrast the different techniques; and therefore gives a better insight of these techniques.

For each technique, we first present, if possible, the *de fecto* statistical model behind the scene. Apart from it, we also list the following aspects of the algorithm.

- Optimization perspective: as many ML technique can be rephrased as an optimization problem, we therefore list its optimization equivalence;
- Model learning algorithm: known as model estimation from statistics community, according to Wasserman [2];
- (Closed-form) solution of the model: also known as estimator from the statistics perspective; if possible, the sampling distribution of the estimator is also presented.
- Regularization: how the model can be regularized to strike the variance-bias trade-off;
- Regression surface/Decision boundary: this criteria helps visually understand what the final model looks like in the original feature space; for classification problem, it is the decision boundary of the classifiers.

1.1 Regression Model

A general regression problem can be collectively modelled as follows: assume data features X_i are fixed (i.e. not random), while the response variables Y_i are generated as some unknown transformation f of X_i plus some random disturbance ε_i:

$$Y_i = f(X_i) + \varepsilon_i, \tag{1}$$

where $\boldsymbol{\varepsilon}$ follows some probability distribution $P(\boldsymbol{\varepsilon})$. Usually, we assume ε_i are i.i.d samples from a zero-mean Gaussian with some constant variance, σ^2, i.e.

$$\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

The response variables $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$ then becomes a random vector with a multivariate Gaussian likelihood (induced by $\boldsymbol{\varepsilon}$) with mean and variance

$$\mathbb{E}\left[\boldsymbol{Y}|\boldsymbol{X}\right] = f(\boldsymbol{X}) = [f(X_1), \ldots, f(X_n)]', \mathrm{Var}[\boldsymbol{Y}|\boldsymbol{X}] = \sigma^2\boldsymbol{I}$$

respectively. Therefore, the log likelihood of this regression model becomes

$$\begin{aligned} \log P(\boldsymbol{Y}|f, \sigma^2, \boldsymbol{X}) = \log \mathcal{N}(\boldsymbol{Y}; f(X), \sigma^2\boldsymbol{I}) &= \sum_{i=1}^{n} \log \mathcal{N}(Y_i; f(X_i), \sigma^2) \\ &= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n}(Y_i - f(X_i))^2}_{\text{squared error loss}}. \end{aligned}$$

Note that the likelihood actually defines a squared error loss function with respect to f, assuming σ^2 is known. And the squared error loss function is

$$L(f) = \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

In fact, in most statistical models, likelihood function serves the same purpose as a loss function, and both of them measures some distance between the assumed model and the observed data. In this case, the loss function simply measures the squared euclidean distance between n-dimensional vectors $\boldsymbol{Y}$ and $f(\boldsymbol{X})$, which is also the kernel of a (spherical) Gaussian distribution. Based on the distance measure, we obtain the optimized solution w.r.t the loss function and likelihood function:

$$f_{ML} = \arg\max_f \log P(\boldsymbol{Y}|f, \sigma^2, \boldsymbol{X}); \;\; f_{loss} = \arg\max_f L(f) \tag{2}$$

the model for linear regression Linear regression is a specific case of the general regression model with the additional assumption on f: it assumes

$$f(X_i) = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p} = \boldsymbol{\beta}^T X, \tag{3}$$

then f is uniquely defined by its parameter $\boldsymbol{\beta}$. Plugging the definition of f into the loss and likelihood function, we obtain the corresponding functions for the linear regression model. The closed form solution to the loss function and likelihood are called least squared estimator and maximum likelihood estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{Y}.$$

References

1. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer Series in Statistics (2009), http://dx.doi.org/10.1007/978-0-387-84858-7
2. Wasserman, L.: All of statistics: a concise course in statistical inference. Springer Science & Business Media (2013)