Statistical learning - Vinciotti (2022)

Stefano Cretti
telegram: @StefanoCretti
Github: https://github.com/StefanoCretti/StatisticalLearning.git

March 6, 2022

Contents

I Introduction **2**

1 General course information **3**
1.1 Textbooks . . . 3
1.2 Assessment . . . 3
1.3 Topics . . . 3

II Statistical learning **5**

2 What is statistical learning? **6**
2.1 Definition of statistical learning . . . 6
2.2 Why estimate f? . . . 6
2.2.1 Prediction . . . 6
2.2.2 Inference . . . 7
2.3 How to extimate f? . . . 7
2.3.1 Parametric methods . . . 8
2.3.2 Non-parametric methods . . . 8
2.3.3 Parametric vs non-parametric methods . . . 8
2.4 Bias-variance trade-off . . . 8

Part I

Introduction

Chapter 1

General course information

1.1 Textbooks

- James et al (2021), Introduction to statistical learning in R, 2nd edition. (Book that is used as guidline for the course, but further concepts will be added during the lectures)
- Hastie et al (2001), Elements of statistical learning. (More advanced book for those who want to study the subject more in depth)

1.2 Assessment

- Three homework tasks during the course (Uploaded on moodle, two weeks of time for each, more practical and mainly focused on applying methods to some data. If done well they will add 2 points to the written exam score)
- Final written exam (More theoretical but still connected to the practical part, for instance by commenting on analysis output)

1.3 Topics

- Linear regression (Gauss, 1800) (Assumed to be already known from Statistical Learning 1)
- Linear discriminant analysis, LDA (Fisher, 1936) (Later extended to quadratic discriminant analysis, QDA)
- Logistic regression (1940s)
- Generalized linear models (Nelder and Wedderburn, 1972)
- Classification and regression trees (Breiman and Freidman, 1980s) (First introduction of computer intensive methods)
- Machine learning (1990s): support vector machines, neural networks/deep learning, unsupervised learning (clustering, PCA)
- Individual methods: theory, details, implementation ...

- General concept: model selection, inference, prediction ...

Part II

Statistical learning

Chapter 2

What is statistical learning?

2.1 Definition of statistical learning

In general, a statistical learning problem can be formalized as follows:

- Y : response/dependent/outcome variable
- $\underline{X} = (X_1, \ldots, X_p)$: vector of predictors/features/independent variables/covariates

We assume that there is a relationship between Y and $\underline{X}$, which can be written as:

$$Y = f(\underline{X}) + \varepsilon$$

Where:

- $f(\underline{X})$ is the deterministic (but unknown) function of the vector $\underline{X} = (X_1, \ldots, X_p)$
- ε is the error (stochastic part), for which we assume the following properties:
  - $E[\varepsilon] = 0$ (Its expected value is zero)
  - $\varepsilon \perp \underline{X}$ (It is independent from $\underline{X}$)

Therefore, the expression **statistical learning** encompasses different methods to estimate $f(\underline{X})$.

2.2 Why estimate f?

There are two main reasons to estimate f, those two being **prediction** and **inference**.

2.2.1 Prediction

Predict Y when we only have observations about $\underline{X}$. Since $E[\varepsilon] = 0$, we usually take:

$$\hat{Y} = \hat{f}(\underline{X})$$

With $\hat{f}$ being our estimate of f.

If this is the only reason to estimate f, then $\hat{f}$ can be a black-box method (deep learning). The accuracy of $\hat{Y}$ as a predictor of Y can be described calculating the expected MSE (mean squared error):

$$\begin{aligned}
E[(Y-\hat{f}(\underline{X}))^2 \mid \underline{X}=\underline{x}] = & && \text{where } \hat{f} \text{ is a fixed known function}\\
= E[(f(\underline{X})+\varepsilon-\hat{f}(\underline{X}))^2] & && \text{since } Y=f(\underline{X})+\varepsilon\\
= E[(f(\underline{X})-\hat{f}(\underline{X}))+\varepsilon)^2] & && \text{rearranging}\\
= E[(f(\underline{X})-\hat{f}(\underline{X}))^2+\varepsilon^2+2\varepsilon(f(\underline{X})-\hat{f}(\underline{X}))] & && \text{solving the square}\\
= E[(f(\underline{X})-\hat{f}(\underline{X}))^2]+E[\varepsilon^2]+2E[\varepsilon(f(\underline{X})-\hat{f}(\underline{X}))] & && \text{separating the expectations}
\end{aligned}$$

Furthermore, since we know that:

$$\begin{aligned}
Var(\varepsilon) &= E[(\varepsilon-E(\varepsilon))^2] && \text{formal definition of variance}\\
&= E[\varepsilon^2]-(E[\varepsilon])^2 && \text{definition generally used during calculation}\\
&= E[\varepsilon^2] && \text{since } E[\varepsilon]=0
\end{aligned}$$

Thus we get:

$$\begin{aligned}
E[(Y-\hat{f}(\underline{X}))^2 \mid \underline{X}=\underline{x}] = & && \\
= E[(f(\underline{X})-\hat{f}(\underline{X}))^2]+Var(\varepsilon)+2E[\varepsilon(f(\underline{X})-\hat{f}(\underline{X}))] & && \text{substituting } E[\varepsilon^2]=Var(\varepsilon)\\
= (f(\underline{X})-\hat{f}(\underline{X}))^2+Var(\varepsilon)+2(f(\underline{X})-\hat{f}(\underline{X}))E[\varepsilon] & && \text{since } f(\underline{X})-\hat{f}(\underline{X}) \text{ is a constant}\\
= (f(\underline{X})-\hat{f}(\underline{X}))^2+Var(\varepsilon) & && \text{since } E[\varepsilon]=0
\end{aligned}$$

With:

- $(f(\underline{X})-\hat{f}(\underline{X}))^2$ being the **reducible error**. The model choice can increase or reduce this value, hence it is mostly controllable.
- $Var(\varepsilon)$ being the **irreducible error**. This value depends on the innate randomness present in the data, hence you can only try and minimize it by deciding which variables to use in your prediction (but it will never be zero otherwise you would have a deterministic situation).

2.2.2 Inference

Inference is used when you want to understand the relation between Y and $\underline{X}$ (and not just be able to make predictions). Namely, inference answers questions such as:

- Which predictors/factors are most associated with the response?
- What is the relationship between Y and X_j?

2.3 How to extimate f?

Given some **training data** $(\underline{x}_i, y_i)$, $i=1,\dots,n$, where $\underline{x}_i=(x_{i1},\dots,x_{ip})^t$ is the vector of observations of unit i while y_i is the response for unit i, broadly speaking there are two types of methods to estimate f: **parametric methods** and **non-parametric methods**.

2.3.1 Parametric methods

In order to use parametric methods, we make an assumption about the functional form of $f(\underline{X})$, that assumption being that the form of the function depends on some parameters (which we can estimate). An example of parametric method is **linear regression**, which implies that $f(\underline{X})$ is in the form:

$$f(\underline{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Given this assumption, statistical learning becomes **fitting** (or training) the model on the data, which means estimating the parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ such that:

$$\hat{f}(\underline{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

The main disadvantage of these methods is that they may be **too restrictive**.
Notice that linear models are linear in the parameters, not in the predictors, hence:

- $f(X_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \varepsilon$ (polynomial regression) is a linear model.
- $f(X_1) = \beta_0 X_1^{\beta_1}$ is not a linear model.

2.3.2 Non-parametric methods

Non-parametric methods do not make any explicit assumption on the function form of $f(\underline{X})$. These methods want to estimate f by getting as close as possible to the data, without being too *rough or wiggly* (basically **overfitting**).

2.3.3 Parametric vs non-parametric methods

Despite parametric methods being more restrictive than non-parametric ones, we might still choose to adopt the former for the sake of interpretability and generalizability outside of the training data.

2.4 Bias-variance trade-off

Assume you have some data pairs in the form (x_i, y_i), $i = 1, \ldots, n$; you can then define the estimate function $\hat{f}(x)$ as a linear model with up to $n-1$ parameters (more parameters give the same result as $n-1$ parameters) and an intercept value. You could thus use, for instance, the models:

1 parameter	$f(x) = \beta_0 + \beta_1 x + \varepsilon$
2 parameters	$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
$\vdots$	$\vdots$
$n-1$ parameters	$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{n-1} x^{n-1} + \varepsilon$

You could choose the model with the highest number of parameters; this model would explain all the variance of the training data ($R^2 = 1$) yet it would be very complex and perform badly with new data points. On the other hand a model with a lower number of parameters would explain less of the variance of the training data, yet it could perform better with new data points.

Keeping in mind that Y is random and that $\hat{f}$ is a random variable estimated from the data, when using the general formula to determine how well a model performs at generic $\underline{X}$, we notice:

$$
\begin{aligned}
&E[(Y-\hat{f}(\underline{X}))^2 \,|\, \underline{X}=\underline{x}] = \\
&\quad = E[(f(\underline{X})+\varepsilon-\hat{f}(\underline{X}))^2] && \text{since } Y = f(\underline{X})+\varepsilon \\
&\quad = E[(f(\underline{X})+\varepsilon-\hat{f}(\underline{X})+E[\hat{f}(\underline{X})]-E[\hat{f}(\underline{X})])^2] && \text{since } E[\hat{f}(\underline{X})]-E[\hat{f}(\underline{X})]=0 \\
&\quad = E[((f(\underline{X})-E[\hat{f}(\underline{X})])+\varepsilon+(E[\hat{f}(\underline{X})]-\hat{f}(\underline{X})))^2] && \text{grouping} \\
&\quad = E[(f(\underline{X})-E[\hat{f}(\underline{X})])^2]+E[\varepsilon^2]+E[(E[\hat{f}(\underline{X})]-\hat{f}(\underline{X}))^2]+ \\
&\qquad + 2E[(f(\underline{X})-E[\hat{f}(\underline{X})])\,\varepsilon]+ \\
&\qquad + 2E[(f(\underline{X})-E[\hat{f}(\underline{X})])(E[\hat{f}(\underline{X})]-\hat{f}(\underline{X}))]+ && \text{solving the square and} \\
&\qquad + 2E[(E[\hat{f}(\underline{X})]-\hat{f}(\underline{X}))\,\varepsilon] && \text{dividing the expectations}
\end{aligned}
$$

But notice that:

$$
\begin{aligned}
E[(f(\underline{X})-E[\hat{f}(\underline{X})])\,\varepsilon] &= E[\varepsilon](f(\underline{X})-E[\hat{f}(\underline{X})]) && \text{since } f(\underline{X})-E[\hat{f}(\underline{X})] \text{ is constant} \\
&= 0 && \text{since } E[\varepsilon]=0
\end{aligned}
$$

$$
\begin{aligned}
&E[(f(\underline{X})-E[\hat{f}(\underline{X})])(E[\hat{f}(\underline{X})]-\hat{f}(\underline{X}))] = \\
&\qquad = (f(\underline{X})-E[\hat{f}(\underline{X})])E[E[\hat{f}(\underline{X})]-\hat{f}(\underline{X})] && \text{since } f(\underline{X})-E[\hat{f}(\underline{X})] \text{ is constant} \\
&\qquad = (f(\underline{X})-E[\hat{f}(\underline{X})])E[\hat{f}(\underline{X})-\hat{f}(\underline{X})] && \text{since } \hat{f}(\underline{X}) \text{ is a constant} \\
&\qquad = (f(\underline{X})-E[\hat{f}(\underline{X})])E[0] \\
&\qquad = 0
\end{aligned}
$$

$$
\begin{aligned}
E[(E[\hat{f}(\underline{X})]-\hat{f}(\underline{X}))\,\varepsilon] &= E[\varepsilon]E[E[\hat{f}(\underline{X})]-\hat{f}(\underline{X})] && \text{since } \varepsilon \perp \underline{X} \implies E[\varepsilon\underline{X}]=E[\varepsilon]E[\underline{X}] \\
&= 0 && \text{since } E[\varepsilon]=0
\end{aligned}
$$

Therefore we can simplify as:

$$
\begin{aligned}
E[(Y-\hat{f}(\underline{X}))^2 \,|\, \underline{X}=\underline{x}] &= E[(f(\underline{X})-E[\hat{f}(\underline{X})])^2] && +E[\varepsilon^2] && +E[(E[\hat{f}(\underline{X})]-\hat{f}(\underline{X}))^2] \\
&= f(\underline{X})-E[\hat{f}(\underline{X})])^2 && +Var(\varepsilon) && +Var(\hat{f}(\underline{X}))
\end{aligned}
$$

Where:

- $f(\underline{X})-E[\hat{f}(\underline{X})])^2$ is the bias
- $Var(\varepsilon)$ is the irreducible error
- $Var(\hat{f}(\underline{X}))$ is the variance of the model

Bias On the other hand, bias refers to the error that is introduced by approximating a problem, which may be extremely complicated, by a much simpler model. If an estimated model performs well on the training data but it does not perform well on new data, the estimated model has high bias. If an estimated model performs well on multiple data sets, the estimated model has low bias. High bias means that an estimated model is far from the real model.
Variance refers to the amount by which would change if we estimated it using a different training data set. Ideally, the $\hat{f}$ calculated through different datasets should be not significantly different. The more flexible models are generally more influenced by a change of dataset.
In order to minimize the expected error, we need to achieve low bias and low variance. In practice, one needs to find a good trade-off between bias and variance, since reducing one often involves increasing the other. In general:

- A **simpe model** has high bias (it is far from the real model) and low variance (when fitting using different training data you get similar estimated parameters).
- A **complex model** has low bias and high variance.

Both in parametric and non-parametric methods, you generally have at least one **tuning parameter**, which is a parameter that can be tweaked (for instance the degree of the polinomial) in order to choose the balance between bias and variance.