

Machine Learning (Semester 1 2025)

Statistical Learning

Nina Hernitschek

Centro de Astronomía CITEVA
Universidad de Antofagasta

April 29, 2025

Motivation

Building on our statistical knowledge so far, we will see a variety of methods to **better understand data**.

With the massive data sets we nowadays often deal with, this part is essential for drawing conclusions from our data.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

recap: Goal of Statistical Inference

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

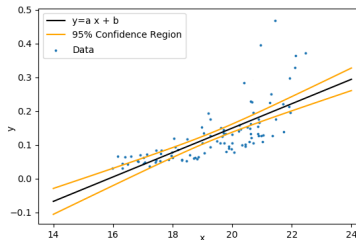
Assessing
Model
Accuracy

Summary &
Outlook

Statistical inference is about **drawing conclusions from data**, specifically determining the properties of a population by data sampling.

Three examples of inference are:

1. What is the best estimate for a (set of) model parameter(s)?



recap: Goal of Statistical Inference

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

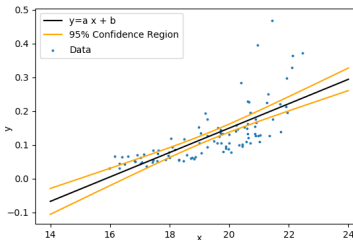
Assessing
Model
Accuracy

Summary &
Outlook

Statistical inference is about **drawing conclusions from data**, specifically determining the properties of a population by data sampling.

Three examples of inference are:

1. What is the best estimate for a (set of) model parameter(s)?
2. How confident we are about our result?



recap: Goal of Statistical Inference

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

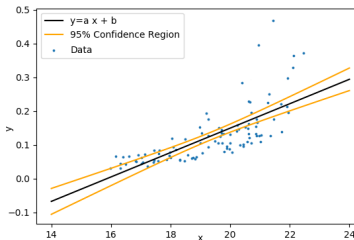
Assessing
Model
Accuracy

Summary &
Outlook

Statistical inference is about **drawing conclusions from data**, specifically determining the properties of a population by data sampling.

Three examples of inference are:

1. What is the best estimate for a (set of) model parameter(s)?
2. How confident we are about our result?
3. Are the data consistent with a particular model/hypothesis?



recap: Some Terminology

We study the properties of some **population** by measuring **samples** from that population. The population doesn't have to refer to different objects.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

recap: Some Terminology

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

We study the properties of some **population** by measuring **samples** from that population. The population doesn't have to refer to different objects.

example: E.g., we may be (re)measuring the position of an object at rest; the population is the distribution of (an infinite number of) measurements smeared by the uncertainty, and the sample are the measurement we've actually taken.

subsequent brightness measurements of a star:

ra dec hjd mag magErr filter

347.66112 -7.39883 2458277.96036 20.083 0.135 g

347.66111 -7.39883 2458280.94526 20.49 0.163 g

347.66111 -7.39881 2458283.94197 19.822 0.116 g

347.66113 -7.39883 2458289.93875 20.361 0.155 g

347.66111 -7.39883 2458377.75728 20.103 0.137 r

347.66111 -7.39883 2458380.84366 20.291 0.151 r

347.66111 -7.39883 2458430.66968 20.471 0.162 r

recap: Some Terminology

Motivation

Frequentist vs.
Bayesian
Inference

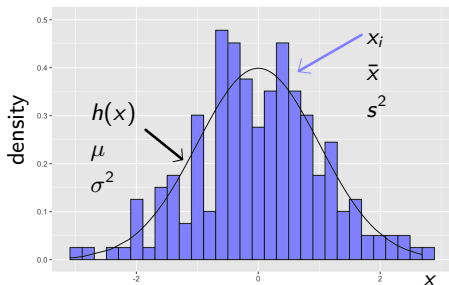
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

A **statistic** is any function of the sample. For example, the sample mean is a statistic. But also something like *the value of the first measurement* is also a statistic.

To conclude something about the population from the sample, we use **estimators**. An estimator is a statistic, a rule for calculating an estimate of a given quantity based on observed data.



recap: Some Terminology

There are **point estimators** and **interval estimators**. The point estimators yield single-valued results (example: the position of an object), while with an interval estimator, the result would be a range of plausible values (example: confidence interval for the position of an object).

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

recap: Some Terminology

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy


Summary &
Outlook

There are **point estimators** and **interval estimators**. The point estimators yield single-valued results (example: the position of an object), while with an interval estimator, the result would be a range of plausible values (example: confidence interval for the position of an object).

Measurements have **uncertainties** (not errors) and we need to account for these (sometimes they are unknown).

Frequentist vs. Bayesian Statistical Inference

There are two major statistical paradigms that address the statistical inference questions:



**classical (frequentist)
paradigm**

Bayesian paradigm

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Frequentist vs. Bayesian Statistical Inference

There are two major statistical paradigms that address the statistical inference questions:

Key differences

classical (frequentist) paradigm

Bayesian paradigm

Definition of probabilities:

relative frequency of events over repeated experimental trials

probabilities quantify our subjective belief about experimental outcomes, model parameters, or models

Quantifying uncertainty:

confidence levels describe the distribution of the measured parameter from the data around the true value

credible regions derived from posterior probability distributions encode our belief in model parameters

Motivation

Frequentist vs. Bayesian Inference

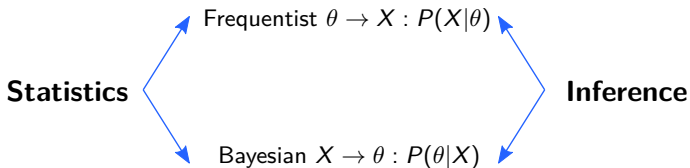
Statistical Learning

Assessing Model Accuracy

Summary & Outlook

Frequentist vs. Bayesian Statistical Inference

we can summarize this as



Motivation

Frequentist vs.
Bayesian
Inference

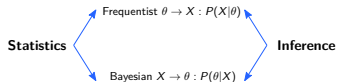
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Frequentist vs. Bayesian Statistical Inference

an example:



A person takes an IQ test (which does not give the “real” IQ but is a way to estimate it, and a possible range of values).

Motivation

Frequentist vs.
Bayesian
Inference

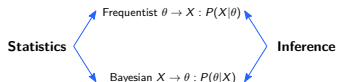
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Frequentist vs. Bayesian Statistical Inference

an example:



A person takes an IQ test (which does not give the “real” IQ but is a way to estimate it, and a possible range of values).

For a **frequentist**, the best estimator is the **average** of many test results. So, if 5 IQ tests were taken and the sample mean is of 160, then that would be the estimator of that candidate’s true IQ.

Motivation

Frequentist vs.
Bayesian
Inference

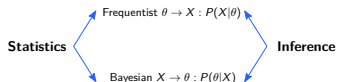
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Frequentist vs. Bayesian Statistical Inference

an example:



A person takes an IQ test (which does not give the “real” IQ but is a way to estimate it, and a possible range of values).

For a **frequentist**, the best estimator is the **average** of many test results. So, if 5 IQ tests were taken and the sample mean is of 160, then that would be the estimator of that candidate’s true IQ.

A **Bayesian** would say: *IQ tests are calibrated with a mean of 100, standard deviation of 15 points* and use this as a **prior** information. The Bayesian estimate of that candidate’s person thus would be not 160, but rather 148, or more specifically that $p(141.3 \leq \mu \leq 154.7 \mid \bar{x} = 160) = 0.683$.

Motivation

Frequentist vs.
Bayesian
Inference

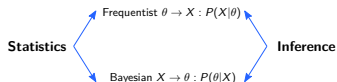
Statistical
Learning

Assessing
Model
Accuracy

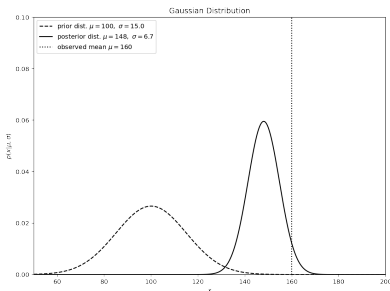
Summary &
Outlook

Frequentist vs. Bayesian Statistical Inference

an example:



A **Bayesian** would say: *IQ tests are calibrated with a mean of 100, standard deviation of 15 points* and use this as a **prior** information. The Bayesian estimate of that candidate's person thus would be not 160, but rather 148, or more specifically that $p(141.3 \leq \mu \leq 154.7 | \bar{x} = 160) = 0.683$.



we will see an astronomy example later in `lecture_2.ipynb`

Bayesian Statistical Inference

With Bayesian statistics, probability expresses a **degree of belief in an event**. This method is different from the frequentist methodology in a number of ways. One of the big differences is that probability actually expresses the chance of an event happening.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Statistical Inference

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

With Bayesian statistics, probability expresses a **degree of belief in an event**. This method is different from the frequentist methodology in a number of ways. One of the big differences is that probability actually expresses the chance of an event happening.

The Bayesian concept of probability is also more **conditional**. In addition to experiment data to predict probabilities (as in the frequentist case), it also uses **prior knowledge**.

Bayesian Statistical Inference

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

With Bayesian statistics, probability expresses a **degree of belief in an event**. This method is different from the frequentist methodology in a number of ways. One of the big differences is that probability actually expresses the chance of an event happening.

The Bayesian concept of probability is also more **conditional**. In addition to experiment data to predict probabilities (as in the frequentist case), it also uses **prior knowledge**.

example: Measuring the flux of a star.

Bayesian Statistical Inference

example: Measuring the flux of a star.
repeated measurements of flux (from nonvariable star) lead to different values due to the statistical error of the astronomical instrument

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Statistical Inference

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

example: Measuring the flux of a star.

repeated measurements of flux (from nonvariable star) lead to different values due to the statistical error of the astronomical instrument

Frequentist: probability only has meaning in terms of a limiting case of repeated measurements

Limit of large numbers: the frequency of any given value indicates the probability of measuring that value.

⇒ probabilities fundamentally related to frequencies of events

Bayesian Statistical Inference

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

example: Measuring the flux of a star.

repeated measurements of flux (from nonvariable star) lead to different values due to the statistical error of the astronomical instrument

Frequentist: probability only has meaning in terms of a limiting case of repeated measurements

Limit of large numbers: the frequency of any given value indicates the probability of measuring that value.

⇒ probabilities fundamentally related to frequencies of events

Bayesian:

the concept of probability is extended to cover degrees of certainty about statements.

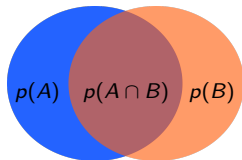
Bayesian approach claims to measure the flux F with a probability $P(F)$: probability as statement of the knowledge of the measurement outcome.

⇒ probabilities fundamentally related to our own knowledge about an event, the **prior**

Bayes' Rule

recap from lecture 1:

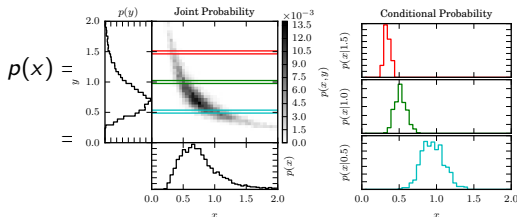
If we have two events, A and B , the possible combinations are:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

$$p(A \cap B) = p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

We then had seen that the **marginal probability** (projecting onto one axis) is defined as



Bayes' Rule

Since $p(x|y)p(y) = p(y|x)p(x)$ we can write that

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

which gives

Bayes' Rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

which in words says that

the (conditional) probability of y given x is just the (conditional) probability of x given y times the (marginal) probability of y divided by the (marginal) probability of x, where the latter is just the integral of the numerator.

The Bayesian Method

The Essence of the Bayesian Method:

- **Probability statements** are not limited to data, but can be made **for model parameters** and models themselves.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

The Bayesian Method

The Essence of the Bayesian Method:

- **Probability statements** are not limited to data, but can be made **for model parameters** and models themselves.
- Inferences are made by producing probability density functions (pdfs); most notably, model parameters are treated as random variables.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

The Bayesian Method

The Essence of the Bayesian Method:

- **Probability statements** are not limited to data, but can be made **for model parameters** and models themselves.
- Inferences are made by producing probability density functions (pdfs); most notably, model parameters are treated as random variables.
- These pdfs represent our belief spread in what the model parameters are. They have nothing to do with outcomes of repeated experiments (although the shape of resulting distributions can often coincide).

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Statistical Inference

frequentist statistical inference:

We calculated a **likelihood** $p(D | M)$.

Bayesian statistical inference:

We instead evaluate the **posterior probability** taking into account prior information and the likelihood.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Statistical Inference

frequentist statistical inference:

We calculated a **likelihood** $p(D | M)$.

Bayesian statistical inference:

We instead evaluate the **posterior probability** taking into account prior information and the likelihood.

We've seen that Bayes' Rule is:

$$p(M | D) = \frac{p(D | M) p(M)}{p(D)},$$

with data D and model $M = M(\theta)$.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Statistical Inference

frequentist statistical inference:

We calculated a **likelihood** $p(D | M)$.

Bayesian statistical inference:

We instead evaluate the **posterior probability** taking into account prior information and the likelihood.

We've seen that Bayes' Rule is:

The diagram shows the equation for Bayes' Rule:
$$p(M | D) = \frac{p(D | M) p(M)}{p(D)}$$
 Blue arrows point from labels to parts of the equation: 'posterior' points to $p(M | D)$, 'likelihood' points to $p(D | M)$, 'prior' points to $p(M)$, and 'evidence' points to $p(D)$.

with data D and model $M = M(\theta)$.

prior probability

How probable are the possible values of θ in nature?

likelihood

ties the model to the data:
how likely is the data given θ ?

posterior probability

distribution is updated with information from the data:
what is the probability of different θ values given data and model?

Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

1. formulate the likelihood, $p(D | M, \theta, I)$

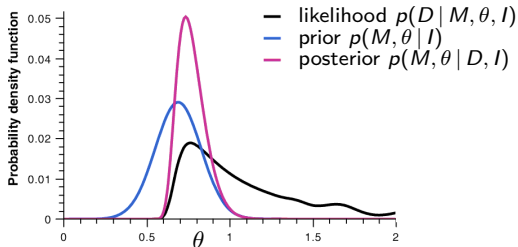
Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

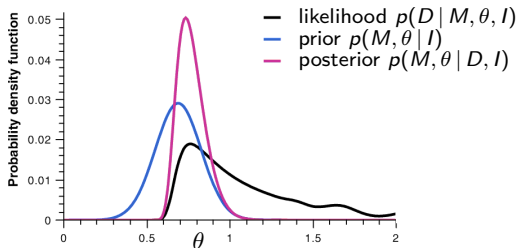
Summary &
Outlook



Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

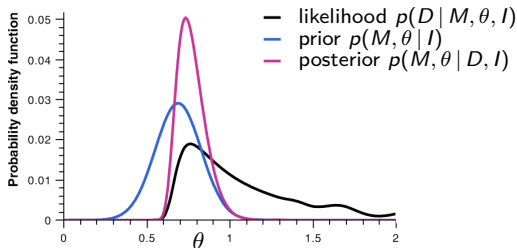
1. formulate the likelihood, $p(D | M, \theta, I)$
2. chose a prior, $p(M, \theta | I)$, which incorporates other information beyond the data in D



Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

1. formulate the likelihood, $p(D | M, \theta, I)$
2. chose a prior, $p(M, \theta | I)$, which incorporates other information beyond the data in D
3. determine the posterior pdf, $p(M, \theta | D, I)$



Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

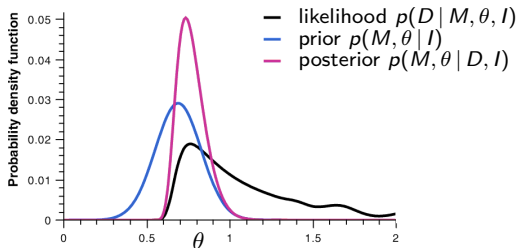
Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

1. formulate the likelihood, $p(D | M, \theta, I)$
2. chose a prior, $p(M, \theta | I)$, which incorporates other information beyond the data in D
3. determine the posterior pdf, $p(M, \theta | D, I)$
4. search for the model parameters that maximize $p(M, \theta | D, I)$



Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

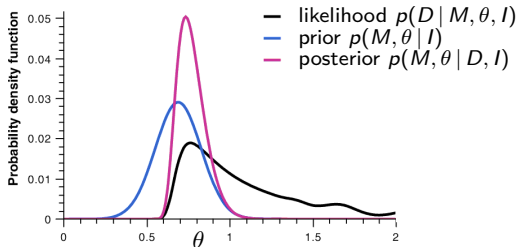
Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

1. formulate the likelihood, $p(D | M, \theta, I)$
2. chose a prior, $p(M, \theta | I)$, which incorporates other information beyond the data in D
3. determine the posterior pdf, $p(M, \theta | D, I)$
4. search for the model parameters that maximize $p(M, \theta | D, I)$
5. quantify the uncertainty of the model parameter estimates



Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

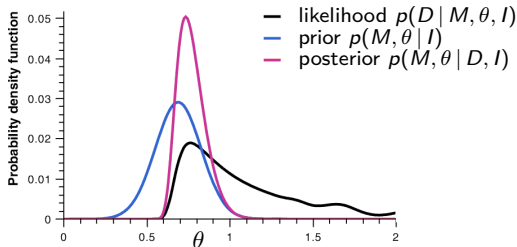
Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

1. formulate the likelihood, $p(D | M, \theta, I)$
2. chose a prior, $p(M, \theta | I)$, which incorporates other information beyond the data in D
3. determine the posterior pdf, $p(M, \theta | D, I)$
4. search for the model parameters that maximize $p(M, \theta | D, I)$
5. quantify the uncertainty of the model parameter estimates
6. perform model selection to find best description of the data



Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

What is Statistical Learning?

Statistical learning refers to methodologies for **understanding data**.

⇒ allows to **uncover hidden correlation patterns**

Motivation

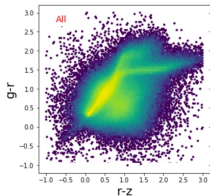
Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

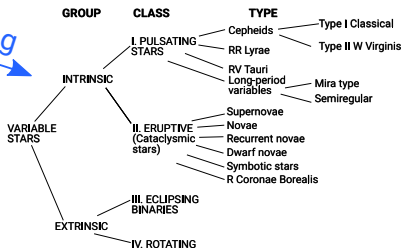
Summary &
Outlook

parameter space of
measurements



machine learning

parameter space of astrophysical
objects



Unsupervised vs. Supervised Learning

Those methods can be distinguished as follows:

unsupervised learning or “learning without labels”

Clustering:

Find groups that are not defined a priori based on measurements

⇒ members of the same cluster are “close” in some sense

vs.

supervised learning or “learning with labels”

Classification:

Use a priori group labels in analysis to assign new observations to a particular group or class

Regression:

instead of having training data with discrete labels, the “truth” is a continuous property and we are trying to predict the values of that property for the test data

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Unsupervised vs. Supervised Learning

supervised learning or “learning without labels”

Classification:

Use a priori group labels in analysis to assign new observations to a particular group or class

Regression:

instead of having training data with discrete labels, the “truth” is a continuous property and we are trying to predict the values of that property for the test data

example:

The task of determining whether an object is a star, a galaxy, or a quasar is a classification problem: the label is from three distinct categories. On the other hand, we might wish to estimate the age of an object based on such observations: this would be a regression problem, because the label (age) is a continuous quantity.

Motivation

Frequentist vs.
Bayesian
Inference

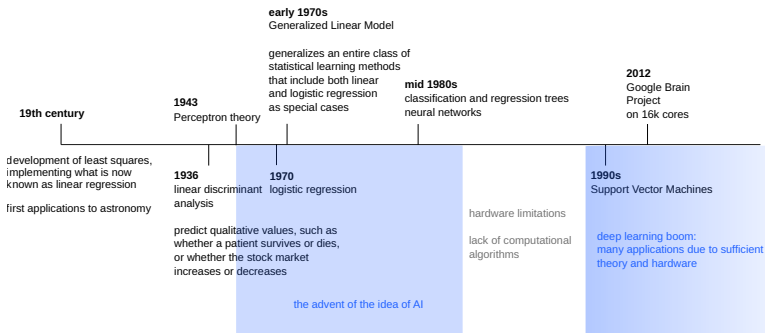
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

A Brief History of Statistical Learning

A fairly new field - but building on concepts rooting in the 19th century:



Motivation

Frequentist vs. Bayesian Inference

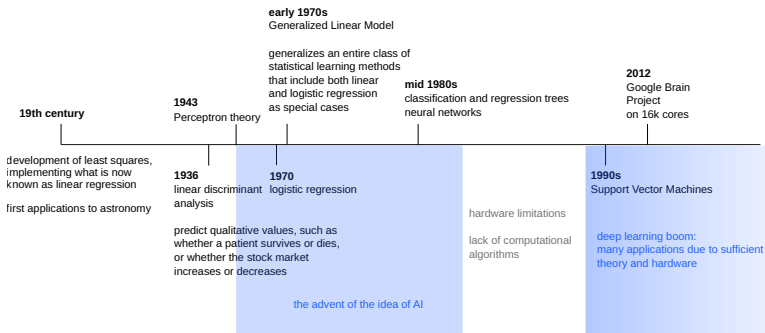
Statistical Learning

Assessing Model Accuracy

Summary & Outlook

A Brief History of Statistical Learning

A fairly new field - but building on concepts rooting in the 19th century:



In recent years, progress in statistical learning has been marked by the increasing availability of powerful and relatively user-friendly software, such as programming libraries for the open-source Python ecosystem.

Notation

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

The input **variables** (predictors, independent variables, features) are typically denoted using the symbol X , with a subscript to distinguish them.

Example: When we ask which kind of advertisement works best for a product, X_1 might be the TV budget, X_2 the radio budget, and X_3 the internet budget.

The **output variable** is often called the response or dependent variable, and is typically denoted using the symbol Y . In this example, the output variable is sales.

Example: Y might be product sales.

More generally, suppose that we observe a quantitative response Y and we have n different predictors, X_1, X_2, \dots, X_n . We assume that there is some relationship between Y and $X = (X_1, X_2, \dots, X_n)$ which can be written in the very general form $Y = f(X) + \epsilon$.

Here f is some fixed but unknown function of X_1, X_2, \dots, X_n representing the systematic information that X provides about Y , and ϵ is a random error term, which is independent of X and has mean zero.

Another Use Case

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

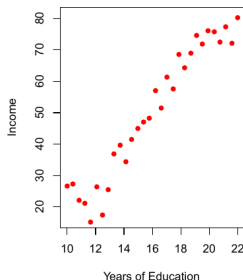
Assessing
Model
Accuracy

Summary &
Outlook

As another use case (outside of astronomy), we consider this example from *An Introduction to Statistical Learning: with Applications in Python*.

Witten, Hastie, Tibshirani; Springer (<https://www.statlearning.com/>)

The data set consists of income vs. years of education for 30 individuals. The plot of income vs. years suggests that one might be able to predict income using years of education.



source: Fig. 2.2 from
<https://www.statlearning.com/>

Another Use Case

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

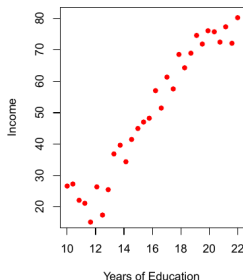
Assessing
Model
Accuracy

Summary &
Outlook

As another use case (outside of astronomy), we consider this example from *An Introduction to Statistical Learning: with Applications in Python*.

Witten, Hastie, Tibshirani; Springer (<https://www.statlearning.com/>)

The data set consists of income vs. years of education for 30 individuals. The plot of income vs. years suggests that one might be able to predict income using years of education.



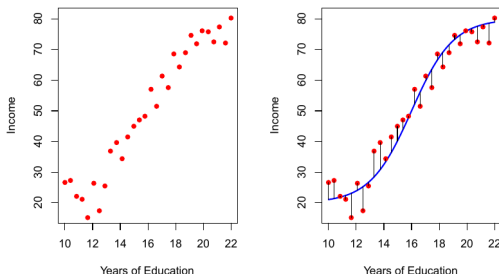
source: Fig. 2.2 from
<https://www.statlearning.com/>



The unknown function f connecting the input variable to the output variable must be estimated based on the observed points.

Another Use Case

As this data set is *simulated*, f is known and is shown by the blue curve in the right-hand panel:



In this plot, the blue curve is the underlying f .

Vertical lines represent the error terms ϵ . Note that some errors are positive and some are negative. Their overall mean is zero.

Motivation

Frequentist vs.
Bayesian
Inference

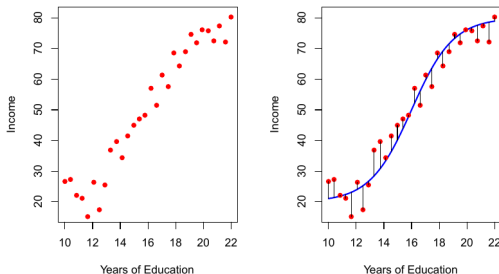
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Another Use Case

As this data set is *simulated*, f is known and is shown by the blue curve in the right-hand panel:



In this plot, the blue curve is the underlying f .

Vertical lines represent the error terms ϵ . Note that some errors are positive and some are negative. Their overall mean is zero.



Statistical learning refers to a set of approaches for estimating f .

Why Estimate f ?

There are two main reasons why we want to estimate f : prediction and inference.

Motivation

Frequentist vs.
Bayesian
Inference

**Statistical
Learning**

Assessing
Model
Accuracy

Summary &
Outlook

Prediction

Often, we have a set of inputs X available, but the output Y cannot be easily obtained.

Based on our assumption of the error term averaging 0, we can predict Y :

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

where \hat{f} represents our estimate for f , and \hat{Y} represents the resulting prediction for Y .

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Prediction

Often, we have a set of inputs X available, but the output Y cannot be easily obtained.

Based on our assumption of the error term averaging 0, we can predict Y :

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

where \hat{f} represents our estimate for f , and \hat{Y} represents the resulting prediction for Y .

example:

Suppose that X_1, \dots, X_n are characteristics of a patient's blood sample, and Y is a variable encoding the patient's risk for a severe reaction to a medication. It is natural to try to predict Y using X , to avoid giving at-risk patients the medication.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Prediction

The accuracy of \hat{Y} as a prediction for Y depends on two quantities: the **reducible error** and the **irreducible error**.

Motivation

Frequentist vs.
Bayesian
Inference

**Statistical
Learning**

Assessing
Model
Accuracy

Summary &
Outlook

Prediction

The accuracy of \hat{Y} as a prediction for Y depends on two quantities: the **reducible error** and the **irreducible error**.

\hat{f} will not be a perfect estimate for f - this inaccuracy will introduce the **reducible error**; we can potentially improve the accuracy of \hat{f} by finding a better f .

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Prediction

The accuracy of \hat{Y} as a prediction for Y depends on two quantities: the **reducible error** and the **irreducible error**.

\hat{f} will not be a perfect estimate for f - this inaccuracy will introduce the **reducible error**; we can potentially improve the accuracy of \hat{f} by finding a better f .

However, even with doing that, our prediction is not free from error. This is because Y is also a function of ϵ , which cannot be predicted by X . This is the **irreducible error**.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Prediction

Consider a given estimate \hat{f} and a set of predictors X , which yields the prediction $\hat{Y} = \hat{f}(X)$.

We assume for the moment that both \hat{f} and X are fixed: now variability only comes from ϵ .

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Prediction

Consider a given estimate \hat{f} and a set of predictors X , which yields the prediction $\hat{Y} = \hat{f}(X)$.

We assume for the moment that both \hat{f} and X are fixed: now variability only comes from ϵ .

Then, it is easy to show that

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(x) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible}} \end{aligned}$$

where $E(Y - \hat{Y})^2$ represents the **expectation value** of the squared difference between the predicted and actual value of Y , and $\text{Var}(\epsilon)$ represents the **variance** associated with the error term ϵ .

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Inference

We are often interested in understanding the **association** between Y and X_1, \dots, X_n . For this we want to estimate f without necessarily making predictions for Y .

Here, we need to know the exact form of \hat{f} .

In this setting, one might want to answer the following questions:

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Inference

We are often interested in understanding the **association** between Y and X_1, \dots, X_n . For this we want to estimate f without necessarily making predictions for Y .

Here, we need to know the exact form of \hat{f} .

In this setting, one might want to answer the following questions:

- Which predictors are associated with the response? Often only a small fraction of the available predictors are strongly associated with Y . Identifying them among a large set of possible variables can be extremely useful, depending on the application.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Inference

We are often interested in understanding the **association** between Y and X_1, \dots, X_n . For this we want to estimate f without necessarily making predictions for Y .

Here, we need to know the exact form of \hat{f} .

In this setting, one might want to answer the following questions:

- Which predictors are associated with the response? Often only a small fraction of the available predictors are strongly associated with Y . Identifying them among a large set of possible variables can be extremely useful, depending on the application.
- What is the relationship between the response and each predictor? Some predictors may have a positive relationship with Y , in the sense that larger values of the predictor are associated with larger values of Y . Other predictors may have the opposite relationship.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Inference

We are often interested in understanding the **association** between Y and X_1, \dots, X_n . For this we want to estimate f without necessarily making predictions for Y .

Here, we need to know the exact form of \hat{f} .

In this setting, one might want to answer the following questions:

- Which predictors are associated with the response? Often only a small fraction of the available predictors are strongly associated with Y . Identifying them among a large set of possible variables can be extremely useful, depending on the application.
- What is the relationship between the response and each predictor? Some predictors may have a positive relationship with Y , in the sense that larger values of the predictor are associated with larger values of Y . Other predictors may have the opposite relationship.
- Depending on the complexity of f , the relationship between the response and a given predictor may also depend on other predictors.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Inference

We are often interested in understanding the **association** between Y and X_1, \dots, X_n . For this we want to estimate f without necessarily making predictions for Y .

Here, we need to know the exact form of \hat{f} .

In this setting, one might want to answer the following questions:

- Which predictors are associated with the response? Often only a small fraction of the available predictors are strongly associated with Y . Identifying them among a large set of possible variables can be extremely useful, depending on the application.
- What is the relationship between the response and each predictor? Some predictors may have a positive relationship with Y , in the sense that larger values of the predictor are associated with larger values of Y . Other predictors may have the opposite relationship.
- Depending on the complexity of f , the relationship between the response and a given predictor may also depend on other predictors.
- Can the relationship be adequately modeled with a linear equation, or is the relationship more complicated?

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Prediction vs. Inference

Some modeling could be conducted both with prediction and inference.

example: For real estate prices, one might want to relate home values to inputs such as neighborhood, distance to downtown, crime rates and so forth.

In this case one might be interested in the association between each individual input variable and housing price - for instance, by how much will the price increase for having ocean view? This is an **inference problem**.

Alternatively, one may simply be interested in predicting the value of a home given its characteristics to estimate whether this house is house under- or over-valued? This is a **prediction problem**.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Prediction vs. Inference

Depending on whether our goal is prediction, inference, or a combination of the two, **different methods** for estimating f may be appropriate.

For example, **linear models** allow for relatively simple and interpretable inference, but may not yield as accurate predictions as some other approaches. In contrast, some of the highly non-linear approaches that we see later in this course can provide quite accurate predictions for Y , but at the expense of a less **interpretable** model.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

How to Estimate f ?

Throughout this course, we will explore many linear and non-linear approaches for estimating f .

We provide an overview of their shared characteristics here before we look at the individual methods.

Motivation

Frequentist vs.
Bayesian
Inference

**Statistical
Learning**

Assessing
Model
Accuracy

Summary &
Outlook

How to Estimate f ?

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Throughout this course, we will explore many linear and non-linear approaches for estimating f .

We provide an overview of their shared characteristics here before we look at the individual methods.

In the following, we assume that we have observed a set of n data points. These observations are called the **training data**: we will use these observations to train our method how to estimate f .

Let x_{ij} represent the value of the j th predictor, or input, for observation i , where $i = 1, 2, \dots, m$ and $j = 1, \dots, n$. Correspondingly, let y_i represent the response variable for the i th observation.

Then our training data consist of $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$.

How to Estimate f ?

Throughout this course, we will explore many linear and non-linear approaches for estimating f .

We provide an overview of their shared characteristics here before we look at the individual methods.

In the following, we assume that we have observed a set of n data points. These observations are called the **training data**: we will use these observations to train our method how to estimate f .

Let x_{ij} represent the value of the j th predictor, or input, for observation i , where $i = 1, 2, \dots, m$ and $j = 1, \dots, n$. Correspondingly, let y_i represent the response variable for the i th observation.

Then our training data consist of $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$.

Our **goal** is to apply a statistical learning method to the training data in order to estimate the unknown function f : We want to find a function \hat{f} such that $Y \sim \hat{f}(X)$ for any observation (X, Y) .

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Parametric Methods

Most statistical learning methods can be characterized as either parametric or non-parametric.

Parameteric methods involve a two-step model-based approach:

1. Make an assumption about the functional form of f . For example, assume that f is a linear function of X :

$$f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

This greatly simplifies the problem as now only the $n + 1$ coefficients need to be estimated.

2. After selecting a model, the training data are needed to fit or train the model. In the case of the linear model, we need to estimate the parameters $\beta_0, \beta_1, \dots, \beta_n$ such that

$$Y \sim \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

The most common approach is called **least squares**.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Parametric Methods

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

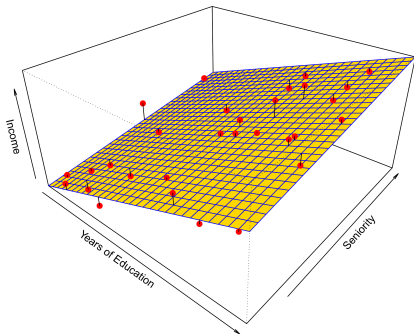
Assessing
Model
Accuracy

Summary &
Outlook

The **advantage** of the parametric approach is that it is much easier to estimate a set of parameters than to fit an entirely arbitrary function.

The **disadvantage** of the parametric approach is that we have a fixed functional form which usually doesn't match the true form of f .

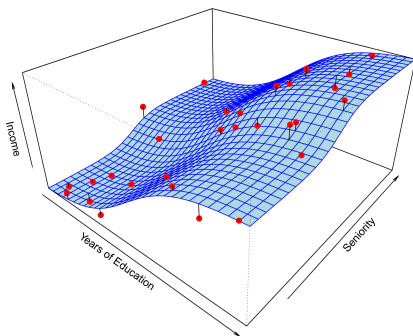
In the following, we see data to which we have fit a linear model of the form $\hat{f} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$:



Linear model fit to data.
The observations are shown in red.
The yellow plane indicates the
least-squares fit to the data.
Source: Fig. 2.4 from
<https://www.statlearning.com/>

Parametric Methods

As the data were simulated by drawing from an underlying f we know we can compare how well the fit works:



The blue surface represents the true underlying relationship.

The observations are shown in red.

Source: Fig. 2.3 from

<https://www.statlearning.com/>

We can see that the linear fit misses some features of the functional form of f , i.e. the true f has some curvature that cannot be captured with a linear fit.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Over- and Underfitting

The statistical learning approach tells us what the best-fitting model parameters are, but **not how good the fit actually is**.

If the model isn't well suited for the data, then we should not expect a good fit.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

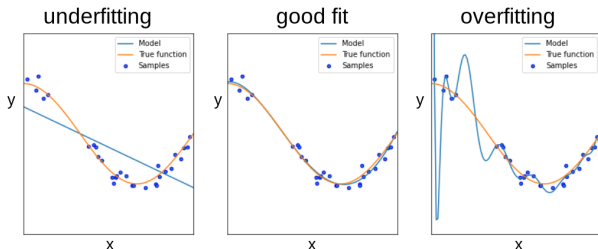
Over- and Underfitting

The statistical learning approach tells us what the best-fitting model parameters are, but **not how good the fit actually is**.

If the model isn't well suited for the data, then we should not expect a good fit.

example:

N points drawn from a linear distribution can always be fitted perfectly with an $N - 1$ order polynomial - which won't help to predict future measurements



Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Non-Parametric Methods

In contrast to parametric methods for which the functional form of f is predetermined, **non-parametric methods** do not make explicit assumptions about the functional form.

Instead they try to estimate the functional form of f without too much over- or underfitting.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Non-Parametric Methods

In contrast to parametric methods for which the functional form of f is predetermined, **non-parametric methods** do not make explicit assumptions about the functional form.

Instead they try to estimate the functional form of f without too much over- or underfitting.

The major **advantage** of non-parametric methods over parametric methods:

The potential to accurately fit a wider range of possible shapes for f .

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Non-Parametric Methods

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

In contrast to parametric methods for which the functional form of f is predetermined, **non-parametric methods** do not make explicit assumptions about the functional form.

Instead they try to estimate the functional form of f without too much over- or underfitting.

The major **advantage** of non-parametric methods over parametric methods:

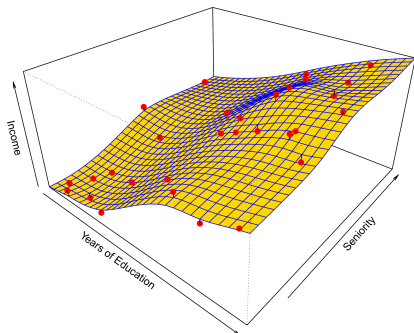
The potential to accurately fit a wider range of possible shapes for f .

The major **disadvantage** of non-parametric methods:

As not reducing the problem of estimating f down to a small set of parameters, a much larger number of observations is required in order to obtain an accurate estimate for f .

Non-Parametric Methods

In the following **example**, we again fit the same data set as shown before, but now with a non-parameteric approach.



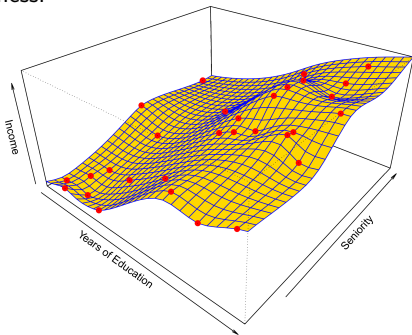
A non-parametric model (thin-plate spline¹, high smooth parameter) is used to estimate f for fitting the data.

Source: Fig. 2.5 from <https://www.statlearning.com/>

¹ *splines* are functions defined piecewise by polynomials

Non-Parametric Methods

The thin-plate spline has a smoothness parameter that must be selected. The following figure shows the same fit as above but using a lower level of smoothness.



Again a thin-plate spline fit, but with a lower level of smoothness.
Source: Fig. 2.6 from <https://www.statlearning.com/>

Motivation

Frequentist vs.
Bayesian
Inference

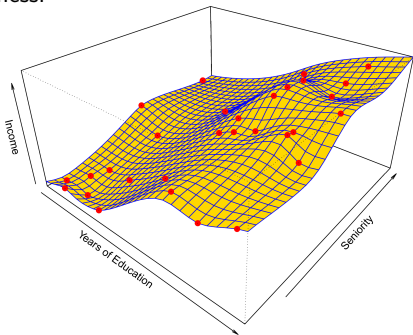
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Non-Parametric Methods

The thin-plate spline has a smoothness parameter that must be selected. The following figure shows the same fit as above but using a lower level of smoothness.



Again a thin-plate spline fit, but with a lower level of smoothness.

Source: Fig. 2.6 from <https://www.statlearning.com/>

We see that the fit follows the data perfectly. Is this a **good result**?

Motivation

Frequentist vs.
Bayesian
Inference

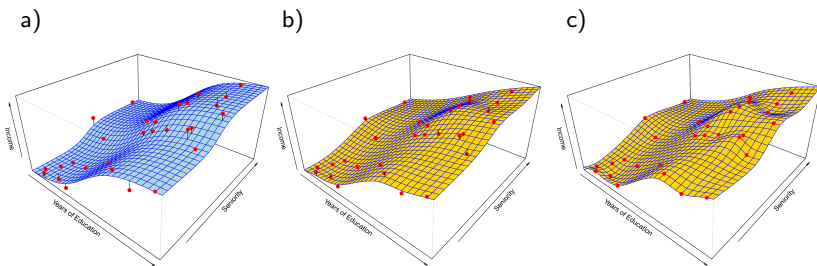
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Non-Parametric Methods

Here we know the true underlying function f as these are simulated data. We can compare our fits to that.



a) The true function f . b) Thin-plate spline, high smooth parameter.
c) Thin-plate spline, low smooth parameter.

Motivation

Frequentist vs.
Bayesian
Inference

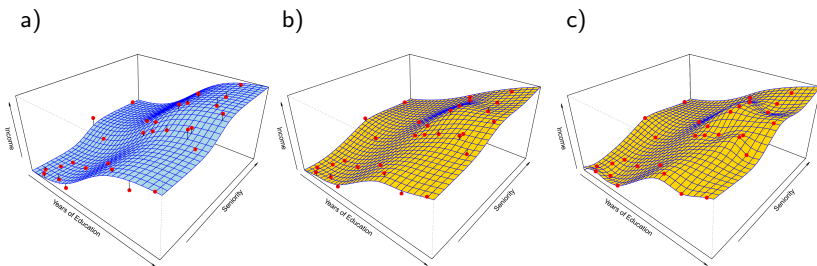
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Non-Parametric Methods

Here we know the true underlying function f as these are simulated data. We can compare our fits to that.



- a) The true function f . b) Thin-plate spline, high smooth parameter.
c) Thin-plate spline, low smooth parameter.

c) is an example for **overfitting**!

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

The Trade-Off Between Prediction Accuracy and Model Interpretability

We have seen that some approaches offer less, some offer more flexibility in estimating f .

For example, linear regression can only generate linear functions (lines, planes...), thus being a relatively rigid approach.

Other methods, such as splines, are considerably more flexible because they can generate a much wider range of possible shapes to estimate f .

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

The Trade-Off Between Prediction Accuracy and Model Interpretability

We have seen that some approaches offer less, some offer more flexibility in estimating f .

For example, linear regression can only generate linear functions (lines, planes...), thus being a relatively rigid approach.

Other methods, such as splines, are considerably more flexible because they can generate a much wider range of possible shapes to estimate f .



Why use a more restrictive method instead of a very flexible approach?

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

The Trade-Off Between Prediction Accuracy and Model Interpretability

There are several reasons that we might **prefer a more restrictive model**.

If we are mainly interested in inference, then restrictive models are much more interpretable.

For instance, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between Y and X_1, X_2, \dots, X_p . In contrast, very flexible approaches, such as splines, can lead to such complicated estimates of f that it is difficult to understand how **any individual predictor** is associated with the response.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Assessing Model Accuracy

Why is it necessary to introduce so many different statistical learning approaches, rather than just a single best method?

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

**Assessing
Model
Accuracy**

Summary &
Outlook

Assessing Model Accuracy

Why is it necessary to introduce so many different statistical learning approaches, rather than just a single best method?

There is not a single method that's best.

On a particular data set, one specific method may work best, but some other method may work better on a similar but different data set.

We thus must decide based on the given data set which method produces the best results. Selecting the best approach can be one of the most **challenging** parts of performing statistical learning in practice.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Model Comparison

For the **Bayesian** approach to model comparison, we start with Bayes' Theorem,

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$p(M, \theta | D, I) = \frac{p(D | M, \theta, I) \times p(M, \theta | I)}{p(D | I)},$$

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Model Comparison

For the **Bayesian** approach to model comparison, we start with Bayes' Theorem,

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$p(M, \theta | D, I) = \frac{p(D | M, \theta, I) \times p(M, \theta | I)}{p(D | I)},$$

and marginalize over model parameter space θ to obtain the **probability of model M** given data D and prior information I :

$$\begin{aligned} p(M | D, I) &\equiv \int p(M, \theta | D, I) d\theta \\ &= \int \frac{p(D | M, \theta, I) p(M, \theta | I)}{p(D | I)} d\theta \\ &= \frac{p(M | I)}{p(D | I)} \int p(D | M, \theta, I) p(\theta | M, I) d\theta \end{aligned}$$

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Model Comparison

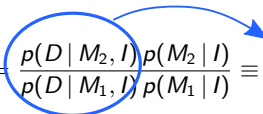
To then determine which of two models is better we compute the ratio of the posterior probabilities or the **odds ratio** as

$$O_{21} \equiv \frac{p(M_2|D, I)}{p(M_1|D, I)}.$$

The posterior probability that the model M is correct given data D (a number between 0 and 1) is

$$p(M|D, I) = \frac{p(D|M, I)p(M|I)}{p(D|I)}.$$

We finally get for the odds ratio:

$$O_{21} = \frac{p(D|M_2, I)p(M_2|I)}{p(D|M_1, I)p(M_1|I)} \equiv B_{21} \frac{p(M_2|I)}{p(M_1|I)},$$


where B_{21} is called the **Bayes factor**.

Motivation

Frequentist vs.
Bayesian
Inference

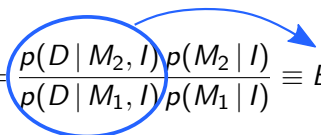
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Model Comparison

We finally get for the odds ratio:

$$O_{21} = \frac{p(D | M_2, I) p(M_2 | I)}{p(D | M_1, I) p(M_1 | I)} \equiv B_{21} \frac{p(M_2 | I)}{p(M_1 | I)},$$


where B_{21} is called the **Bayes factor**.

The Bayes factor compares how well the models fit the data.

It is a ratio of data likelihoods averaged over all allowed values of the model parameters. For models fitting the data equally well, decision is made based on the priors.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Model Comparison

We finally get for the odds ratio:

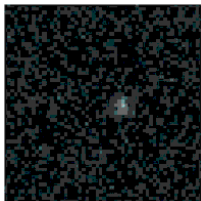
$$O_{21} = \frac{p(D | M_2, I) p(M_2 | I)}{p(D | M_1, I) p(M_1 | I)} \equiv B_{21} \frac{p(M_2 | I)}{p(M_1 | I)},$$

where B_{21} is called the **Bayes factor**.

The Bayes factor compares how well the models fit the data.

It is a ratio of data likelihoods averaged over all allowed values of the model parameters. For models fitting the data equally well, decision is made based on the priors.

example: Consider a noisy image of a source which is equally likely to be a star or a galaxy. The posterior probability that the source is a star will greatly depend on whether we are looking at the Galactic plane or not.



Bayesian Model Comparison

We can compute

$$E(M) \equiv p(D | M, I) = \int p(D | M, \theta, I) p(\theta | M, I) d\theta,$$

where $E(M)$ is called the **marginal likelihood for model M** (or **evidence** or *fully marginalized likelihood*). It quantifies the probability that the data D would be observed if the model M were the correct model.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Model Comparison

We can compute

$$E(M) \equiv p(D | M, I) = \int p(D | M, \theta, I) p(\theta | M, I) d\theta,$$

where $E(M)$ is called the **marginal likelihood for model M** (or **evidence** or *fully marginalized likelihood*). It quantifies the probability that the data D would be observed if the model M were the correct model.

The evidence is a weighted average of the likelihood function, where the prior for model parameters acts as the weighting function.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Model Comparison

How do we **interpret** the values of the odds ratio in practice?

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Model Comparison

How do we **interpret** the values of the odds ratio in practice?

Jeffreys (1936, 1961) proposed a scale for interpreting the odds ratio, where $O_{21} > 10$ represents strong evidence in favor of M_2 (M_2 is ten times more probable than M_1), and $O_{21} > 100$ is decisive evidence (M_2 is one hundred times more probable than M_1). When $O_{21} < 3$, the evidence is not worth more than a bare mention.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Bayesian Model Comparison

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

How do we **interpret** the values of the odds ratio in practice?

Jeffreys (1936, 1961) proposed a scale for interpreting the odds ratio, where $O_{21} > 10$ represents strong evidence in favor of M_2 (M_2 is ten times more probable than M_1), and $O_{21} > 100$ is decisive evidence (M_2 is one hundred times more probable than M_1). When $O_{21} < 3$, the evidence is not worth more than a bare mention.

caution:

- These are just definitions of conventions, i.e., a way to give a quantitative meaning to qualitative phrases.
- The odds ratio compares the models, it doesn't tell us about the absolute goodness of fit.
Model A can be $100\times$ better than model B (by the odds ratio or another measure), but still don't fit the data well.

Approximate Bayesian Model Comparison

The full odds ratio can be costly to compute \Rightarrow **approximate methods** that balance between *goodness of fit* and *model complexity*.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Approximate Bayesian Model Comparison

Akaike information criterion (AIC)

$$\text{AIC} \equiv -2 \ln[L_0(M)] + 2k + \frac{2k(k+1)}{N-k-1}.$$

with

k : number of model parameters

$L_0(M)$: maximum value of the likelihood function

The term $\frac{2k(k+1)}{N-k-1}$ is sometimes ignored.

The **preferred model** is the one with the minimum AIC value. AIC rewards goodness of fit (as assessed by the likelihood function), but also includes a penalty for an increasing number of parameters to discourage overfitting.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Approximate Bayesian Model Comparison

Bayesian information criterion (BIC)

The BIC can be derived from the Bayesian odds ratio **by assuming that the likelihood is Gaussian**.

⇒ easier to compute than the odds ratio as it is based on the maximum value of the likelihood, $L_0(M)$, rather than on the integration of the likelihood over the full parameter space (i.e. evidence $E(M)$).

The BIC is for N data points and a model with k parameters:

$$\text{BIC} \equiv -2 \ln[L_0(M)] + k \ln N.$$

The 1st term is equal to the model's χ^2 (under the assumption of normality; note that this is not χ^2 per degree of freedom!).

The 2nd term on the right hand side penalizes complex models relative to simple ones.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Approximate Bayesian Model Comparison

When two models are compared, the model with the smaller BIC/AIC value wins.

If the models are equally successful in describing the data (i.e., they have the same value of $L_0(M)$), then the model with fewer free parameters wins (Occam's razor).

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Approximate Bayesian Model Comparison

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

When two models are compared, the model with the smaller BIC/AIC value wins.

If the models are equally successful in describing the data (i.e., they have the same value of $L_0(M)$), then the model with fewer free parameters wins (Occam's razor).

caution:

Both BIC and AIC are **approximations** and might not be valid if the underlying assumptions (e.g.: Gaussian likelihood) are not met.

⇒ If computationally feasible, compute the odds ratio.

Mean Squared Error

In the **regression** setting, the most commonly-used measure is the mean squared error (MSE), given by

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

where $\hat{f}(x_i)$ is the prediction that \hat{f} gives for the i th observation.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Mean Squared Error

In the **regression** setting, the most commonly-used measure is the mean squared error (MSE), given by

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

where $\hat{f}(x_i)$ is the prediction that \hat{f} gives for the i th observation.

The MSE will be small if the predicted responses are very close to the true responses, and will be large if for some of the observations the prediction differs substantially from the true response.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Mean Squared Error

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

The MSE is computed using the training data to fit the model. For this reason, it should more accurately be referred to as the training MSE.

But: In general, we do not really care how well the method works on the training data. We are interested in the accuracy of the predictions that we obtain when we **apply our method to previously unseen data**.

Suppose we fit our statistical learning method to our training observations $\{(x_1, y_1), \dots, (x_n, y_n)\}$, and we obtain the estimate \hat{f} .

However, we're not interested in whether $\hat{f} \sim y_i$ (training MSE); but we want to know whether $\hat{f}(x_0) \sim y_0$, where (x_0, y_0) is a **previously unseen test observation not within the training set**. This gives us the **test MSE**, and we want to choose the method that gives the lowest test MSE.

Test and Training MSE



How can we select a method that minimizes the test MSE?

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Test and Training MSE



How can we select a method that minimizes the test MSE?

Optimal case: We have a test data set available (that was not used in training). This is usually the case if we have a large set of observations with response. We can then simply split up our observations into a training and test set.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Test and Training MSE



How can we select a method that minimizes the test MSE?

Optimal case: We have a test data set available (that was not used in training). This is usually the case if we have a large set of observations with response. We can then simply split up our observations into a training and test set.

Less optimal case: If we have only a low number of observations with response, we might be tempted to simply selecting a statistical learning method that minimizes the training MSE.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Test and Training MSE



How can we select a method that minimizes the test MSE?

Optimal case: We have a test data set available (that was not used in training). This is usually the case if we have a large set of observations with response. We can then simply split up our observations into a training and test set.

Less optimal case: If we have only a low number of observations with response, we might be tempted to simply selecting a statistical learning method that minimizes the training MSE.

Fundamental problem with this strategy: there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE.

Motivation

Frequentist vs.
Bayesian
Inference

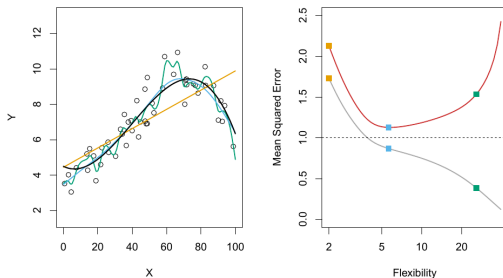
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Test and Training MSE

We illustrate this problem on a simple example.



Source: Fig. 2.9 from <https://www.statlearning.com/>

Left: We have generated observations with the true f given by the black curve. The orange, blue and green curves show spline fits obtained using methods with increasing levels of flexibility. As the level of flexibility (degrees of freedom) increases, the curve fit the data more closely.

Motivation

Frequentist vs.
Bayesian
Inference

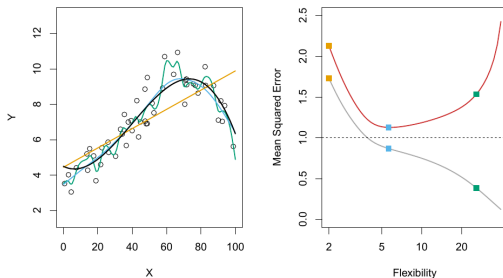
Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Test and Training MSE

We illustrate this problem on a simple example.



Source: Fig. 2.9 from <https://www.statlearning.com/>

Right: The curve gives the average training MSE as a function of flexibility. The training MSE declines monotonically as flexibility increases. The red curve gives the test MSE.

The blue curve minimizes the test MSE. The horizontal dashed line indicates $\text{Var}(\epsilon)$, the irreducible error, which corresponds to the lowest achievable test MSE among all possible methods.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Test and Training MSE

When a given method yields a small training MSE but a large test MSE, we are said to be **overfitting the data**. This happens because the statistical learning algorithm is picking up patterns in the data that are just fluctuations rather than by true properties of the unknown function f . In contrary, the test MSE will be very large when overfitting as the supposed patterns that the method found in the training data simply don't exist in the test data.

Note that regardless of whether or not overfitting has occurred, we almost always expect the training MSE to be smaller than the test MSE because most statistical learning methods either directly or indirectly seek to minimize the training MSE. Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Model Accuracy for Classification

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Suppose now instead of regression, we seek to estimate f on the basis of training observations where now y_1, \dots, y_n are qualitative. A common approach for measuring the accuracy of \hat{f} is the **training error rate**, the **fraction of incorrect classifications** if we apply our estimate \hat{f} to the training observations:

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i).$$

Here \hat{y}_i is the predicted class label for the i th observation using \hat{f} . $I(y_i \neq \hat{y}_i)$ is an indicator variable that equals 1 if $y_i \neq \hat{y}_i$ (classified incorrectly), and 0 if $y_i = \hat{y}_i$ (classified correctly).

Model Accuracy for Classification

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Suppose now instead of regression, we seek to estimate f on the basis of training observations where now y_1, \dots, y_n are qualitative. A common approach for measuring the accuracy of \hat{f} is the **training error rate**, the **fraction of incorrect classifications** if we apply our estimate \hat{f} to the training observations:

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i).$$

Here \hat{y}_i is the predicted class label for the i th observation using \hat{f} . $I(y_i \neq \hat{y}_i)$ is an indicator variable that equals 1 if $y_i \neq \hat{y}_i$ (classified incorrectly), and 0 if $y_i = \hat{y}_i$ (classified correctly).

As in the regression setting, the above gives the training error, whereas we are most interested in the test error rate. The test error rate is given by

$$\text{Ave}(I(y_0 \neq \hat{y}_0)),$$

where \hat{y}_0 is the predicted class label that results from applying the classifier to the test observation with predictor x_0 .

Summary

In this session, we saw an overview about the objectives of statistical learning.

We also saw some applications of classification and regression algorithms, along with methods on how to quantify the quality of fit.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook

Outlook

In the next session, we will focus on the details of **regression algorithms**.

Motivation

Frequentist vs.
Bayesian
Inference

Statistical
Learning

Assessing
Model
Accuracy

Summary &
Outlook