Frequentist Statistics and Bayesian Statistics

Volker Tresp
Summer 2019
Frequentist Statistics
Approach

• Natural science attempts to find regularities and rules in nature

\[ F = ma \]

• The laws are valid under idealized conditions. Example: Fall of a point object without air friction, with velocities much smaller than the speed of light

• There might be measurement errors, but there is an underlying true (simple) dependency

• This motivates the frequentist statistics: *derivation of probabilistic statements under repeatable experiments under identical conditions*
Repeated experiments with an underlying linear dependency

\( m = 1 \text{kg} \)
Basic Terms

• Thus a statistical analysis requires a precise description of the experiment. For example, the details on who gets which medication (randomized?)

• A statistical unit is an object, on which measurements are executed (attributes are registered). Could be a person. A statistical unit defines a row in the data matrix, the attributes define the columns

• The population is the conceptual set of all statistical units about which we want to perform statistical inference. Example: diabetics

• For the analysis, only a sample is available (training data). Often it is assumes that the sample is a random subset of the population
Population

- A population can be finite, infinite, or hypothetic
- Example: all people who vote in an election
Typical Assumption

- The sample $D$ is a random subset of the population
- For each statistical unit $i$ in the sample, we determine the attributes (features) $x_i$
- Assuming a random sample, we can write (in a finite sample, we would assume sampling with replacement) with $P(\cdot)$ known

$$P(D) = P(x_1, \ldots, x_N) = \prod_{i=1}^{N} P(x_i)$$

- The probability that I sample $N$ units with attributes $x_1, \ldots, x_N$ is the product of the probabilities of observing individual units with their individual attributes
Modelling

- $P(x_i)$ is unknown
- Assumption in parametric modelling: The data has been generated by a probability distribution $P_w(x_i)$, which is parameterized by the parameter vector $w$. For example, we might assume a Gaussian distribution with unknown mean but known variance.
- Thus we assume that for at least one parameter vector $w$
  \[ P_w(x_i) \approx P(x_i) \]
- The goal is to estimate the parameter vector
Example: a Person’s Height

- We assume that the height $x_i$ is Gaussian distributed with unknown mean and variance

$$P_w(x_i) = \mathcal{N}(x_i; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right)$$

with $w = (\mu, \sigma)^T$

- Thus we get

$$P_w(x_1, \ldots, x_N) = \prod_{i=1}^{N} P_w(x_i) = \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma^2)$$

$$= \frac{1}{(2\pi \sigma^2)^{N/2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 \right)$$
How can we define and find the best parameters?
Maximum Likelihood

- We consider the probability of the observed data as a function of the parameters. This is the likelihood-function, where we assume that data points were generated independently.

\[ L(w) = P_w(x_1, \ldots, x_N) = \prod_{i=1}^{N} P_w(x_i) \]

- It is often more convenient to work with the log-likelihood,

\[ l(w) = \log L(w) = \sum_{i=1}^{N} \log P_w(x_i) \]

- The maximum likelihood (ML) estimator is given by

\[ \hat{w}_{ml} \doteq \arg \max(l(w)) \]

- This means: in the family of distributions under considerations, the ML estimator is the one which explains the data the best.
• The ML estimators are

\[ \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \]

and

\[ \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \]
ML-Estimator for a Linear Model

• Let’ assume that the true dependency is linear, but we only have available noisy target measurements

\[ y_i = x_i^T w + \epsilon_i \]

• Let’s further assume that the noise is Gaussian distributed

\[ P(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \epsilon_i^2\right) \]

• It follows that

\[ P_w(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - x_i^T w)^2\right) \]

• It is easier to deal with the log

\[ \log P_w(y_i|x_i) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_i - x_i^T w)^2 \]
ML Estimator

- The log-likelihood function is then

\[ l = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T w)^2 \]

- The first term does not depend on \( w \). Under the assumption of independent additive noise, the ML estimator is the same as the LS estimator \( \hat{w}_{ml} = \arg\max(l(w)) = \hat{w}_{LS} \)

Since, \( \hat{w}_{ml} = \arg\max[- \sum_i (y_i - x_i^T w)^2] \) and \( \hat{w}_{ls} = \arg\min[\sum_i (y_i - x_i^T w)^2] \)
Analysis of Estimators

• Certainly the ML estimator makes sense (best fit). But how certain are we about the estimates. Maybe there are parameter values that would give us almost the same likelihood?

• To analyse the ML estimate we do the following thought experiment (see next slide)

• Let $\mu$ be the unknown but fixed parameter

• In addition to the available sample we are able to generate additional samples $D_1, D_2, \ldots D_L$, $L \to \infty$, each of size $N$

• For each of these $D_i$, we estimate the parameter and obtain $\hat{\mu}_i$ (for example, using the ML-estimator)
We analyse the distribution of the estimated parameter.

In the example, we get for the mean person height (with known $\sigma^2$)

\[
P_\mu(\hat{\mu} - \mu) = \mathcal{N}(\hat{\mu} - \mu; 0, \frac{\sigma^2}{N})
\]

The interpretation of probability here is: averaged of all $D_1, D_2, \ldots D_L$

We can calculate this distribution of the difference between estimated and true parameter without knowing any particular data set (although I need $\sigma^2$)

Assuming, we estimate $\hat{\mu}$ from the available sample, we can answer the question: how probable is it to measure $\hat{\mu}$ if the true value is $\mu = 175 cm$?
The frequentist experiment

Distribution of the estimated parameter

\[ P(\hat{\mu} | \mu) \propto \mathcal{N}(\mu, \sigma^2/N) \]
Bias of an Estimator

- The difference between the true parameter and the expected value of the parameter estimate (averaged over many data sets of size $N$) is called the bias

$$Bias(\hat{w}) = E_D(\hat{w}) - w_{true}$$

Here,

$$E_D(\hat{w}) = \lim_{L \to \infty} \frac{1}{L} \sum_{i=1}^{L} \hat{w}_{D_i}$$

In the example, the bias is zero for the mean
The ML-Estimator can be Biased with finite Data

- The ML-estimator can be biased with finite data

\[
\hat{\sigma}^2_{ml} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]

\[
\hat{\sigma}^2_{unbiased} = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]
Variance of an Estimator

- The variance indicates how much an estimator varies around its mean

\[ \text{Var}(\hat{w}) = E_D (\hat{w} - E_D(\hat{w}))^2 \]

\[ \text{Var}(\hat{w}) = \lim_{L \to \infty} \frac{1}{L} \sum_{i=1}^{L} (\hat{w}_{D_i} - E_D(\hat{w}))^2 \]

- In the example: \( \text{Var}(\hat{w}) = \sigma^2 / N \)
For finite $N$

The ML estimator can have a finite bias

For $N \to \infty$

The ML estimator is unbiased
Expected Error

- The expected mean squared error evaluates the deviation of the estimator from the true parameter

\[
MSE(\hat{w}) = E_D (\hat{w} - w_{true})^2
\]

\[
MSE(\hat{w}) = \lim_{L \to \infty} \frac{1}{L} \sum_{i=1}^{L} (\hat{w}_D - w_{true})^2
\]

- The expected mean squared error is the sum of the variance and the square of the bias

\[
MSE(\hat{w}) = Var_D(\hat{w}) + Bias_D^2(\hat{w})
\]
Expected Error (cont’d)

Proof:

\[\text{MSE}(\hat{w}) = E_D (\hat{w} - w_{true})^2 = E_D [(\hat{w} - E_D(\hat{w})) - (w_{true} - E_D(\hat{w}))]^2\]

\[= E_D (\hat{w} - E_D(\hat{w}))^2 + E_D (w_{true} - E_D(\hat{w}))^2\]

\[-2E_D [(\hat{w} - E_D(\hat{w}))(w_{true} - E_D(\hat{w}))] = Var_D(\hat{w}) + \text{Bias}_D(\hat{w})^2 + 0\]

The cross term is zero since

\[E_D [(\hat{w} - E_D(\hat{w}))(w_{true} - E_D(\hat{w}))] =\]

\[(w_{true} - E_D(\hat{w})) E_D(\hat{w} - E_D(\hat{w})) = 0\]
Desirable Properties of Estimators

- An estimator is unbiased, if $Bias(\hat{w}) = 0$
- An estimator is asymptotically unbiased, if $Bias(\hat{w}) = 0$, for $N \to \infty$
- An estimator is MSE consistent, if we have
  \[ MSE(\hat{w})_{N \to \infty} \to 0 \]
- An estimator $\hat{w}$ is MSE-efficient, if
  \[ MSE[\hat{w}] \leq MSE[\tilde{w}] \quad \forall \tilde{w} \]
Properties of the ML-Estimator

The ML-estimator has many desirable properties:

- The ML-estimator is asymptotically \( N \to \infty \) unbiased (although with a finite sample size it might be biased).

- Maybe surprisingly, the ML estimator is asymptotically \( (N \to \infty) \) MSE-efficient among all unbiased estimators.

- Asymptotically, the estimator is Gaussian distributed, even when the noise is not!
Estimating the Variance via Bootstrap

- In particular for complex models it might be difficult to derive the sampling distribution, for example the distribution of the ML parameter estimate.

- Recall that ideally we would have many training sets of the same size available, fit the model, and observe the distribution of the parameter estimates.

- Proxies for the new data sets of the same size $N$ can be generated surprisingly simple: A new data set can be generated by sampling $N$ times from the original data with replacement.
Classical Statistical Inference

- For hypothesis testing and the derivation of error bounds, please consult your favorite statistics book.
Discussion: ML

- The likelihood can be calculated even for complex models (e.g., models with latent variables).

- With the assumption that the data have been generated independently, the log-likelihood is the sum over the log likelihoods of individual data points:

\[ l(w) = \sum_{i=1}^{N} \log P(y_i|w) \]

- Thus a log-likelihood defines a cost function (cross-entropy cost function):

\[ cost_i(w) = - \log P(y_i|w) \]
Discussion: ML (cont’d)

• The necessity to emulate the data generating process leads to interesting problem
  specific models

• A certain problem: One needs to assume that the true model is (approximately) in
  the class of the models under considerations.

• With finite data, the ML estimator can lead to over fitting: more complex models will
  have a higher likelihood

• The frequentist statistics has a strong focus in the analysis of the properties of pa-
  rameter estimates
Violations of IID

- The following decomposition assumes that the data points are independent and identically distributed (IID, or i.i.d.)

\[ L(w) = P_w(x_1, \ldots, x_N) = \prod_{i=1}^{N} P_w(x_i) \]

- Statistical analysis under IID is well studied

- For more complex sampling situations, as in time-series modelling or for graph data, the i.i.d. principle can often not be applied, but one can still define a likelihood for the observed data and one can obtain an ML estimate

- The generalization to new data is often nontrivial and is case specific

- Examples: a social network model where new individuals become known; the generalization of a social network, developed for one university, to another university
Bayesian Statistics
The Bayesian Approach

• In a frequentist setting, the parameters are fixed but unknown and the data are generated by a random process.

• In a Bayesian approach, also the parameters have been generated by a random process.

• This means we need an \textit{a priori} distribution:

\[ P(w) \]

• The we obtain a complete probabilistic model:

\[ P(w)P(D|w) \]

• ... and can calculate the posterior parameter distribution using Bayes’ formula as:

\[ P(w|D) = \frac{P(D|w)P(w)}{P(D)} \]
An Example

• Let’s assume that the height $w$ of all German males of age 20 follows a Gaussian distribution

$$P(w) = \mathcal{N}(w; \mu, \alpha^2)$$

• Now you measure the height of one male German person with some Gaussian measurement noise

$$P(x|w) = \mathcal{N}(x; w, \sigma^2)$$

• An ML estimate of this person’s height, ignoring the prior, would be $\hat{w} = x$, $Var(\hat{w}) = \sigma^2$

• The Bayesian would say that

$$P(w|x) = \frac{P(x|w)P(w)}{P(x)} = \mathcal{N} \left( w; \frac{x + \frac{\sigma^2}{\alpha^2} \mu}{1 + \frac{\sigma^2}{\alpha^2}}, \frac{\sigma^2}{1 + \frac{\sigma^2}{\alpha^2}} \right)$$
Prior Distribution

• In the previous example, even a frequentist might agree that the Bayesian solution makes sense.

• The Bayesian approach goes further: Even if $P(w)$ was not available from prior measurements, the user must specify a $P(w)$ according to the user’s prior belief!

• As if your money (or life) would depend on it!
The Prior

- Does it make sense to assume a personal $P(w)$?

- Cox (1946): If one is willing to assign numbers to one's personal beliefs, then one arrives, under few consistent conditions, at the Bayesian formalism.
The Bayesian Experiment

- In contrast to the frequentist experiment, we only work with the actual data \( D \) and do not need to assume that additional hypothetical data sets can be generated.

- One assume that the true parameter \( \mu \) has been generated from the prior distribution \( P(\mu) \) in one experiment. In the example: \( P(\mu) = \mathcal{N}(\mu; 0, \alpha^2) \)

- The data are generated from \( P(D|\mu) \), in the example \( P(D|\mu) = \prod_i \mathcal{N}(x_i; \mu, \sigma^2) \)

- Applying Bayes’ formula I get the \textit{a posteriori} distribution

\[
P(\mu|D) = \frac{P(D|\mu)P(\mu)}{P(D)} = \mathcal{N} \left( \mu; \frac{\text{mean}}{1 + \frac{\sigma^2}{N\alpha^2}}, \frac{\sigma^2}{1 + \frac{\sigma^2}{N\alpha^2}} \right)
\]

with \textit{mean} = \( \frac{1}{N} \sum_{i=1}^{N} x_i \)
The Bayesian experiment

\[
P(\mu) \propto N(0, \alpha^2)
\]

\[
P(\mu | D) \propto N \left( \frac{\text{mean}}{\frac{\sigma^2}{N\alpha^2}}, \frac{\sigma^2}{N + \frac{\sigma^2}{\alpha^2}} \right)
\]
**Analysis**

- The Bayesian approach gives you the complete a posteriori parameter distribution.

- One can derive a maximum *a posteriori* estimator as,

\[
\hat{w}_{map} = \arg \max (P(w|D))
\]

In the example,

\[
\hat{\mu}_{MAP} = \frac{\text{mean}}{1 + \frac{\sigma^2}{N\alpha^2}}
\]

- Note, that the MAP estimator converges to the ML estimator, for \( N \to \infty \).
Our Favorite Example: Linear Regression

- Assume, that the true dependency is linear but that we only measure noisy target data

\[ y_i = x_i^T w + \epsilon_i \]

We get (same as in the frequentist approach)

\[
P(y_i|w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i^T w)^2\right)
\]
Linear Regression: a priori Assumption

- A convenient \textit{a priori} assumption is that

\[
P(w) = (2\pi \alpha^2)^{-M/2} \exp \left( -\frac{1}{2\alpha^2} \sum_{i=0}^{M-1} w_i^2 \right)
\]

- We give smaller parameters a higher \textit{a priori} probability

- Ockhams razor: simple explanations should be preferred

- We will assume that the hyperparameters $\sigma^2$ and $\alpha^2$ are known. If they are unknown, one can define prior distributions for those. The analysis becomes more involved
Linear Regression: the a posteriori Distribution

- Using the likelihood-function and the prior parameter distribution, we can apply Bayes’ formula and obtain the a posteriori distribution

\[ P(w|D) = \frac{P(w)P(D|w)}{P(D)} \]
Linear Regression: Calculating the a posteriori Distribution

\[ P(w|D) = \frac{P(w)P(D|w)}{P(D)} \propto \exp \left( -\frac{1}{2\alpha^2} \sum_{j=0}^{M-1} w_j^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T w)^2 \right) \]

This can be written as

\[ P(w|D) = \mathcal{N}(w; w_{map}, cov(w|D)) \]

With

\[ w_{map} = \left( X^T X + \frac{\sigma^2}{\alpha^2} I \right)^{-1} X^T y \]

and covariance

\[ cov(w|D) = \sigma^2 \left( X^T X + \frac{\sigma^2}{\alpha^2} I \right)^{-1} \]
Linear Regression: the MAP estimate and the PLS-solution

- The most probable parameter value, after observing the data, is (the maximum a posteriori (MAP) estimate)

\[ \hat{w}_{map} = \arg \max(P(w|D)) = \hat{w}_{Pen} \]

with \( \lambda = \frac{\sigma^2}{\alpha^2} \).

- One sees that despite different experimental assumptions the frequentist ML estimate and the Bayesian MAP estimate are very similar. The ML estimate corresponds to the LS-solution and the MAP estimate corresponds to the PLS solution.
Bayesian Prediction with Linear Regression

An important difference between is prediction. In a frequentist solution one substitutes the parameter estimate \( \hat{y}_i = x_i^T w_{ml} \), and one can calculate the variance in the prediction. In a Bayesian approach one applies the rules of probability and marginalizes (integrates over) the parameters.

With

\[
P(y, w|x, D) = P(w|D)P(y|w, x)
\]

it follows that

\[
P(y|x, D) = \int P(w|D)P(y|w, x)dw
\]
Predictive Distribution for a Linear Model

- The a posteriori predictive distribution becomes

$$P(y|x, D) = \int P(y|w, x) P(w|D)dw$$

$$= \mathcal{N}(y; x^T\hat{w}_{map}, x^T\text{cov}(w|D)x + \sigma^2)$$

and is Gaussian distributed with mean $x^T\hat{w}_{map}$ and variance $x^T\text{cov}(w|D)x + \sigma^2$

- The variance on the prediction considers both the noise on the prediction as well as the uncertainty in the parameters (by integrating over possible values)

- This is an essential advantage of the Bayesian approach: one considers all plausible parameter values and, e.g., one can also consider all local optima in the integral

- This is also the main technical challenge: for the Bayesian solution complex integrals need to be solved or approximated
Discussion: the Bayesian Solution

- Personal belief is formulated as a probability distribution; a mechanism for
- Consistent approach for various kinds of modeling uncertainty
- For basic distributions (Gaussian, Poisson, Dirichlet, ...) which belong to the exponential family of distributions, closed form solutions for the complete Bayesian approach are available!
- For more complex models, a predictive analysis leads to integrals which often cannot be solved analytically
- Special approximations: Monte-Carlo integration, evidence framework
- The simplest approximation is

\[
P(y|x, D) = \int P(y|w, x)P(w|D)dw \approx P(y|x, w_{map})
\]

which means that one uses a MAP point estimate