What we have done so far.

# Prerequisites

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Frequentist vs. Bayesian** |  |
|  | |  |
| There are two main ways to interpret probabilities    The **frequentist way** views probabilities as the result of limiting processes    The **bayesian way** views probabilities as subjective measures of belief | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Example** |  |
|  | |  |
| **Frequentists** believe probabilities can be decomposed into relative frequencies. This means, that even a simple coin toss is the result of a limiting procedure:    This requires the concept of repeatability. As the term above is just an expectation, the have to be independent, identically distributed in order for the law of large numbers to apply.  **Bayesians** have issues with this approach, as not every experiment can be repeated. The probability that somebody is breaking into your apartment tomorrow is something that only happens once and cannot be seen as relative frequency.  Both approaches are sometimes at odds but can be unified if the sequence of random variables is *exchangeable*. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Probabilities is a non-unique measure** |  |
|  | |  |
| For Bayesians, choosing probabilities as a measure of belief is justified by the *Dutch Book theorem*. But using probabilities is not a must (even though convenient).  Other quantifications are possible.  To see a cool video on this, check this out:  https://www.youtube.com/watch?v=GC-l345c1FY | | | |

# Making inferences

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Bayes Theorem** |  |
|  | |  |
| Bayes theorem forms the heart of our inference.  Using it, we are not only able to make inferences about “classical” random variables but also parameters. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Bayes Theorem** |  |
|  | |  |
| Let be random variables. We can then retrieve the conditional probability  as follows: | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Parameters as random variables** |  |
|  | |  |
| If y is a parameter, strict frequentists would have problems interpreting  Bayes’ theorem at first. For them, a parameter is not something that can be experimentally tested. How can you see the probability as relative frequency if is only a theoretical construct?  But if a sequence of is exchangeable, there is a way to unify Frequentists and Bayesians approaches by using “De Finetti’s representation theorem”. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Exchangeable sequences** |  |
|  | |  |
| A sequence is exchangeable, if its probability is invariant under permutations: | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **De Finetti’s thereom** |  |
|  | |  |
| Let be an exchangeable sequence of Bernoulli variables.  Note: The do not have to be independent nor identically distributed.  Then the following holds: | | | |
|  | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Representation theorem Notes** |  |
|  | |  |
| The takeaway: Given an exchangeable sequence, there exists a parameter under which this sequence is conditionally independent and identically distributed.  A prior distribution is merely an opinion about the limit of   This satisfies Frequentists, because they now see that as the result of a limiting process and Bayesians alike, as that shows their parameter distributions give us insights about the data.  For more see a very good answer here:  <https://stats.stackexchange.com/questions/34465/what-is-so-cool-about-de-finettis-representation-theorem>  And the general version (Hewitt-Savage) for non binary can be found here:  http://www.stats.ox.ac.uk/~steffen/teaching/grad/definetti.pdf | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Prior Distributions** |  |
|  | |  |
| Prior distributions give us an educated guess about the distribution of in before we see new data.  As we saw in de Finetti’s representation theorem, they can also be interpreted as an opinion about the limit of  .  When choosing priors, there are different concepts to consider.  When selecting based on **conjugacy**, we try to choose a prior that has the same functional form as the likelihood. This often facilitates easier calculation of the posterior.  When selecting based on **conjugacy**, we try to choose a prior that has the same functional form as the likelihood. This often facilitates easier calculation of the posterior.  #  When selecting based on **invariance principles**, we pick a prior that satisfies certain invariance principles.  Of course, you can also pick priors based on completely arbitrary reasons. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Conjugate Priors** |  |
|  | |  |
| Conjugate priors have the same functional form as the likelihood. This is purely done for computational reasons, as this often facilitates easier calculation of the posterior.  Otherwise, conjugate priors generally have no other advantage. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Example** |  |
|  | |  |
| If our likelihood has the form  it makes sense to also have a normally distributed prior:  We have already seen in the exercises, that completing the square allows us to find a (also normally distributed) form for the posterior  Had we picked a different likelihood, we probably could not have completed the square and might have received a distribution that is intractable. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Invariance Principles** |  |
|  | |  |
| Invariance principles are rules under which our prior distribution should be invariant. These rules are often quite broad (e.g. should not change under reparameterization). | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Example** |  | |
|  | |  | |
| One sometimes unwanted side effect is that prior distributions also change under reparameterizations. Suppose you have a cube. Inside this cube is another hidden cube, whose side length is a number randomly chosen between 1-3 cm. What is the expected volume of the hidden cube?   |  | | --- | | One side is between 1-3 cm. As the side lengths of the hidden cube are uniformly distributed, we have an expected length of .  The volume of the cube is between . If we randomly select the middle, our expected volume is  But how can it be, that our expected side length is but the expected volume ? |   The apparent paradox appears a uniform distribution under transformations is not necessarily uniform. In the above example, . But has density  which is not uniform!  See more here: https://en.wikipedia.org/wiki/Principle\_of\_indifference | | | | |
|  |  | **Jeffrey’s prior** | |  |
|  | |  |
| “Jeffrey’s Prior” gives us a prior distribution that is invariant under reparameterization. | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **(Im)-proper priors** |  |
|  | |  |
| A prior whose product  cannot be normalized is called improper. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Bayesian Updating** |  |
|  | |  |
| If we receive new data, we can make use of proportionality to infer | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Outlook** |  |
|  | |  |
| To learn more about priors, the evidence and likelihoods, visit Machine Learning 2 in summer! | | | |

# Linear Regression

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Linear Regression** |  |
|  | |  |
| Let be random variables. A relationship of the form  is assumed by a linear regression model. is a negligible noise term. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Basis functions** |  |
|  | |  |
| In order to model nonlinearities in the data, we transform the input via basis functions. | | | |

|  |  |
| --- | --- |
| Figure Without basis functions, we can only model  an affine relationship | Figure With basis function, we can also model  nonlinearities in the model. |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Basis vector** |  |
|  | |  |
| To model nonlinearities, we transform our input through basis vectors:  Where and the highest wanted degree wanted. It is common to set the intercept term | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Basis functions** |  |
|  | |  |
| Common basis functions are polynomial and exponential . But anything can be used, even sines! | | | |

|  |  |
| --- | --- |
| Figure Polynomial basis functions still behave like polynomials. Which means the tails diverge on both sides. This may lead to wrong inferences further away from the last samplepoint. | Figure Exponential basis function only give local  predictions. Far away from any data points they  vanish to zero. This behavior is often preferred. |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Design Matrix** |  |
|  | |  |
| The design matrix (often also denoted by ) is defined as  where are samples. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Commonly used assumptions** |  |
|  | |  |
| We generally assume that the sampled are normally distributed and independent from each other, i.e.  This does not always have to be the case (see exercise A.6.3) but is convenient, as this likelihood allows “nice” conjugate priors. | | | |

|  |
| --- |
| *Figure 5 Because we maximize with respect to* ***,*** *you can choose any basis function you like! Depicted is a sinusoidal basis function . It can be argued that the can be removed as sines are non local anyway. A sine as basis function is not stupid as Fourier series can predict a large class of (even semi continuous) functions using combinations of trigonometric functions.* |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Overfitting** |  |
|  | |  |
| Increasing the degrees of freedom allows us to more accurately fit the data.  As additional parameters allow us to fit at least additional samples exactly (underdetermined linear system).  But an exact fit is not always warranted, as the prediction between the  samples might wildly oscillate (see Runge’s Phenomenon for polynomials).  If our predictions become worse even though the likelihood increases, we experience overfitting. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Stone Weierstraß Theorem** |  |
|  | |  |
| The Stone-Weierstraß theorem states, that for every function there exists a polynomial which uniformly converges towards | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Interpretation of Stone Weierstraß** |  |
|  | |  |
| It is tempting to assume, that our polynomial is related to the samples we draw. But this is not the case: Increasing the number of samples does not automatically gives us a polynomial that converges uniformly towards (and thereby avoids overfitting).  To my knowledge, finding such can only be done by knowing the function beforehand (see Bernstein polynomials). | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Example** |  |
|  | |  |
| We have a plethora of tools available to combat overfitting. The most  common one is using a regularizer to minimize the squared error:  where represents a vector norm. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Regularization** |  |
|  | |  |
| The use of a regularizer unifies Bayesians and Frequentists, as  the prior distribution is just an implicit regularizer.  Differences between both philosophies become more apparent, when we look at the role of the evidence in model selection. | | | |

# Not-linear Regression

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Regularization** |  |
|  | |  |
| Even though regression might appear powerful, we know from exercise A4.1 that an optimal MSE estimate is the conditional expectation  Unfortunately, requires us to know the joint densities .  As data is sparse, it might be very hard to estimate them.  Instead of minimizing weights , we can also look in the space of functions for solutions. These are called “kernel based methods”.  It can be shown that both, the weight space and function space view, are theoretically equivalent (Mercer’s theorem).  This approach will be further investigated in Machine Learning 2. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Functions are hard** |  |
|  | |  |
| Looking for functions (even the weaker notion “distribution”) is very hard. Even if conditions are specified (e.g. Lipschitz continuity and boundary values for ordinary differential equations or Neumann conditions for partial differential equations), exact solutions often remains elusive.  The | | | |