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 $P(X = 1) = p$ is the result of a limiting procedure:

$$P(X = 1) = \lim_{n \to \infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = p$$

This requires the concept of repeatability. As the term above is just an expectation, the $X_i's$ 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 x, y be random variables. We can then retrieve the conditional probability as follows:

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

Parameters as random variables

If y= θ 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 $p(\theta)$ as relative frequency if θ is only a theoretical construct? But if a sequence of θ is exchangeable, there is a way to unify Frequentists and Bayesians approaches by using "De Finetti's representation theorem".

Exchangeable sequences

A sequence $X_1, X_2, \dots, X_n$ is exchangeable, if its probability is invariant under permutations:

$$P(X_1, X_2, \dots, X_n) = P\left(X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(n)}\right)$$

De Finetti's thereom

Let $\{X_i\}_{i=1}^{\infty}$ be an exchangeable sequence of Bernoulli variables.
Note: The $X_i's$ do not have to be independent nor identically distributed.

Then the following holds:

$$\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} \to \theta, \qquad mit\ \theta{\sim}\mu(\theta)$$

$$P(X_1 = x_1, \dots, X_n = x_n) = \int_{[0,1]} \theta^s (1-\theta)^{n-s} d\mu(\theta)$$

Representation theorem Notes

The takeaway: Given an exchangeable sequence, there exists a parameter θ under which this sequence is conditionally independent and identically distributed.

A prior distribution $\mu(\theta)$ is merely an opinion about the limit of $\bar{X}_n$. This satisfies Frequentists, because they now see that $p(\theta)$ 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 X_i 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 y in $p(y|x)$ 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 $\bar{X}_i$.

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

$$L(\mu) = c_1 e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

it makes sense to also have a normally distributed prior:

$$p(\mu) = c_2 e^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T \tau_0 \mathbf{I}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}$$

We have already seen in the exercises, that completing the square allows us to find a (also normally distributed) form for the posterior

$$p(\boldsymbol{\mu}|\mathbf{x}) = N\left(\boldsymbol{\mu}; \boldsymbol{\mu}_{post}, \boldsymbol{\Sigma}_{post}\right)$$

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 $E[side] = 2cm$.

The volume of the cube is between $1^3 - 3^3 cm^3$. If we randomly select the middle, our expected volume is $E[volume] = 14cm^3$.

But how can it be, that our expected side length is $2cm$ but the expected volume $14cm^3$?
The apparent paradox appears a uniform distribution under transformations is not necessarily uniform. In the above example, $X \sim unif(1,3)$. But $Y = X^3$ has density

$$f_X(g^{-1}) \cdot \left|\frac{d}{dy} g^{-1}\right| = \frac{1}{6} y^{-\frac{2}{3}}$$

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

$$\int p(x|\theta)p(\theta)dx = p(x)$$

cannot be normalized is called improper.

Bayesian Updating

If we receive new data, we can make use of proportionality to infer

$$p(\theta|x_{new}, x_{old}) \propto p(x_{old}|\theta)p(\theta|x_{old})$$

Outlook

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

Linear Regression

Linear Regression

Let $\mathbf{x}, \mathbf{y}$ be random variables. A relationship of the form

$$\boldsymbol{y} = w_0 + w_1 x_1 + \cdots + w_m x_{\mathrm{m}} + \epsilon$$

is assumed by a linear regression model. ϵ is a negligible noise term.

Basis functions

In order to model nonlinearities in the data, we transform the input x via basis functions.

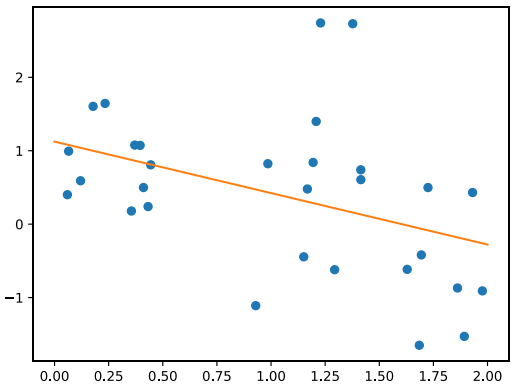

Figure 1 Without basis functions, we can only model an affine relationship $y = w_1 x_1 + w_0$

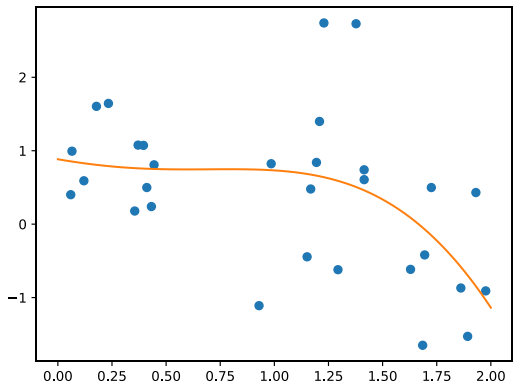

Figure 2 With basis function, we can also model nonlinearities in the model.

Basis vector

To model nonlinearities, we transform our input $\mathbf{x} \in \mathbb{R}^k$ through basis vectors:

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{pmatrix} \phi_0(\mathbf{x}) \\ \dots \\ \phi_m(\mathbf{x}) \end{pmatrix}$$

Where $m = \sum_{i=1}^{d} \binom{k+i-1}{i}$ and d the highest wanted degree wanted. It is common to set the intercept term $\phi_0 = 0$.

Basis functions

Common basis functions are polynomial $\phi_i(x) = x^i$ and exponential $\phi_i(x) = e^{-\frac{1}{2}s(x-\mu_i)^2}$. But anything can be used, even sines!

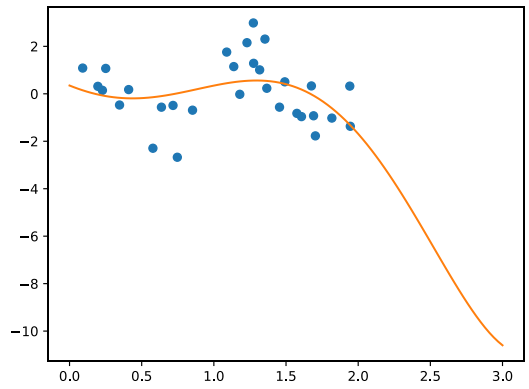

Figure 3 Polynomial basis functions still behave like polynomials. Which means the tails diverge on both sides. This may lead to wrong inferences further away from the last samplepoint.

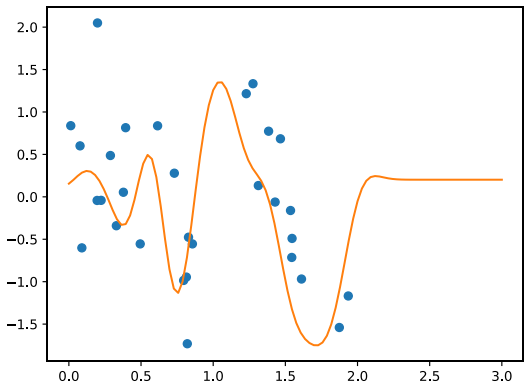

Figure 4 Exponential basis function only give local predictions. Far away from any data points they vanish to zero. This behavior is often preferred.

Design Matrix

The design matrix $\mathbf{\Phi}$ (often also denoted by $\mathbf{X}$) is defined as

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x_1}) & \phi_1(\mathbf{x_1}) & \dots & \phi_m(\mathbf{x_1}) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \dots & \phi_m(\mathbf{x}_2) \\ \dots & \dots & \dots & \dots \\ \phi_0(\mathbf{x}_{n1}) & \phi_1(\mathbf{x_n}) & \dots & \phi_m(\mathbf{x_n}) \end{pmatrix}$$

where $\mathbf{x}_i \in \mathbb{R}^k$ are samples.

Commonly used assumptions

We generally assume that the sampled $(\mathbf{t}_i, \mathbf{x}_i)$ are normally distributed and independent from each other, i.e.

$$(\mathbf{t}_i, \mathbf{x}_i) \sim N(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \sigma_\epsilon^2)$$

This does not always have to be the case (see exercise A.6.3) but is convenient, as this likelihood allows “nice” conjugate priors.

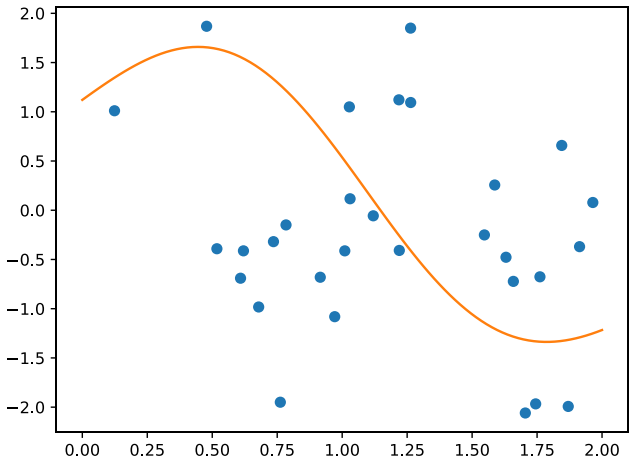

Figure 5 Because we maximize with respect to $\mathbf{w}$*, you can choose any basis function you like! Depicted is a sinusoidal basis function* $\phi_i(x) = sin(\pi(x - \mu_i)^2)$*. It can be argued that the* μ_i *can be removed as sines are non local anyway. A sine as basis function is not stupid as Fourier series can predict a large class of (even semi continuous) functions using combinations of trigonometric functions.*

Overfitting

Increasing the degrees of freedom m allows us to more accurately fit the data.

As k additional parameters allow us to fit at least k 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 C_1 function f there exists a polynomial p which uniformly converges towards f:

$$\lim_{n\to\infty} \sup_{x\in\Omega} |f - p_n| = 0$$

Interpretation of Stone Weierstraß

It is tempting to assume, that our polynomial p_n is related to the n 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 f (and thereby avoids overfitting).

To my knowledge, finding such p_n can only be done by knowing the function f beforehand (see Bernstein polynomials).

Example

We have a plethora of tools available to combat overfitting. The most common one is using a regularizer λ to minimize the squared error:

$$-\frac{1}{2}\sum_{i=1}^{n}\left(t_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\right)^T\boldsymbol{\Sigma}^{-1}\left(t_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\right) + \lambda\|\mathbf{w}\|_p$$

where $\|\mathbf{w}\|_p$ 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 $E[\mathbf{t}|\mathbf{x}]$

Unfortunately, $E[\mathbf{t}|\mathbf{x}]$ requires us to know the joint densities $p(\mathbf{t}, \mathbf{x})$. As data is sparse, it might be very hard to estimate them.

Instead of minimizing weights $\mathbf{w}$, 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