Friedrich-Alexander-Universität
Technische Fakultät

Machine Learning
Data Analytics

Machine Learning for Time Series

(MLTS or MLTS-Deluxe Lectures)

Dr. Dario Zanca

Machine Learning and Data Analytics (MaD) Lab

Friedrich-Alexander-Universität Erlangen-Nürnberg

02.11.2023

Organisational Information

Machine Learning for time series

- 5 ECTS
- Lectures + Exercises

~~**Machine Learning for Time Series (Deluxe)**~~

- ~~7.5 ECTS~~
- ~~Lectures + Exercises + Project~~

Topics overview

- Time series fundamentals and definitions (2 lectures)
- Bayesian Inference (1 lecture) ←
- Gussian processes (2 lectures)
- State space models (2 lectures)
- Autoregressive models (1 lecture)
- Data mining on time series (1 lecture)
- Deep learning on time series (4 lectures)
- Domain adaptation (1 lecture)

Course times

Lectures (online)

A new lecture recording is generally released every Thursday on FAU.TV

Consultation hours by appointment, write to dario.zanca@fau.de

Exercises (online)

Live Zoom Session starting on November 3rd

Recordings from previous editions are available at https://www.fau.tv/course/id/3178

StudOn 2023-2024:
https://www.studon.fau.de/crs5276833.html

Exams and evaluation

Written Exam (5 ECTS)

- 70% from lectures, 30% from exercises
- **On-campus**

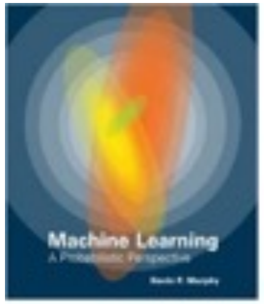
In this lecture…

1. Bayes Theorem
2. Bayesian Model Selection
3. Prior Distributions
4. Linear Regression (Bayesian treatment)

References

Machine learning: A Probabilistic Perspective,

by Kevin Murphy (2012)



Bayesian Inference

Bayes’ Theorem

Bayes' Theorem

Formulation

The **Bayes' Theorem** was formulated by the English philosopher **Thomas Bayes** (1701 – 1761), whose notes were edited and published posthumously by Richard Price.

Didn't publish himself.

Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

this plain version; not so useful.

where A and B are events, and $P(B) \neq 0$.

Portrait from: Terence O'Donnell, *History of Life Insurance in Its Formative Years* (Chicago: American Conservation Co:, 1936), p. 335

Bayes' Theorem

Formulation

Naming Components.

$PP = \frac{L.P}{MP}$

$$\text{Posterior prob} = \frac{\text{Likelihood} \times \text{prior}}{\text{marginal probability}}$$

Posterior probability
The probability of event A occurring given that B is true

Likelihood
The probability of event B occurring given that A is true

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Marginal probability
The probability of observing B without any given conditions

Prior probability
The probability of observing A without any given conditions

Portrait from: Terence O'Donnell, *History of Life Insurance in Its Formative Years* (Chicago: American Conservation Co:, 1936), p. 335

Bayes' Theorem

(Disease)(Correct)
(no disease)(wrong)

Machine Learning Data Analytics FAU

An example. Iterative application of the Bayes' theorem.

Example of application.

$P(B)$ = Testing positive = (Dis.)(test) + (no disease)(not detected).

Initial - 1%.

We know that:

- Disease chance: 1%
- Test accuracy: 95%

$\boldsymbol{A}$: Having the disease
$\boldsymbol{B}$: Testing positive to the disease

$\mathrm{P}(\boldsymbol{B}|\boldsymbol{A})$: Test sensitivity (Likelihood)
$\mathrm{P}(\boldsymbol{B})$: Prob. of a positive test (Marginal)
$\mathrm{P}(\boldsymbol{A})$: Disease chance (Prior)

Having disease - A; first
Testing positive - B; Next.
Marginal Denominator.

First positive test

Now

$$P(A|B) = \frac{.95 \times 0.01}{0.95 \cdot 0.01 + (1 - 0.95)(1 - 0.01)} = 0.161$$

16%.

Second positive test

$$P(A|B) = \frac{.95 \times 0.161}{0.95 \cdot 0.161 + (1 - 0.95)(1 - 0.161)} = 0.785$$

Third positive test

$$P(A|B) = \frac{.95 \times 0.785}{0.95 \cdot 0.785 + (1 - 0.95)(1 - 0.785)} = 0.986$$

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)}$$

If the 2nd test is also positive 78.5% chance disease!

Portrait from: Terence O'Donnell, *History of Life Insurance in Its Formative Years* (Chicago: American Conservation Co:, 1936), p. 335

3 times in a row positive => 98.6%. You have

Bayesian modelling

Example of a linear model

ML perspective.

Let $\mathcal{D}$ denote the **observed data**,

$$\mathcal{D} = \{x^{(n)}, y^{(n)}\}$$

with $x^{(n)} \in \mathcal{R}$ represents the input, and $y^{(n)} \in \mathcal{R}$ represents the output (labels).

(Linear Combination)

Bias ? term.

The **model** is defined as

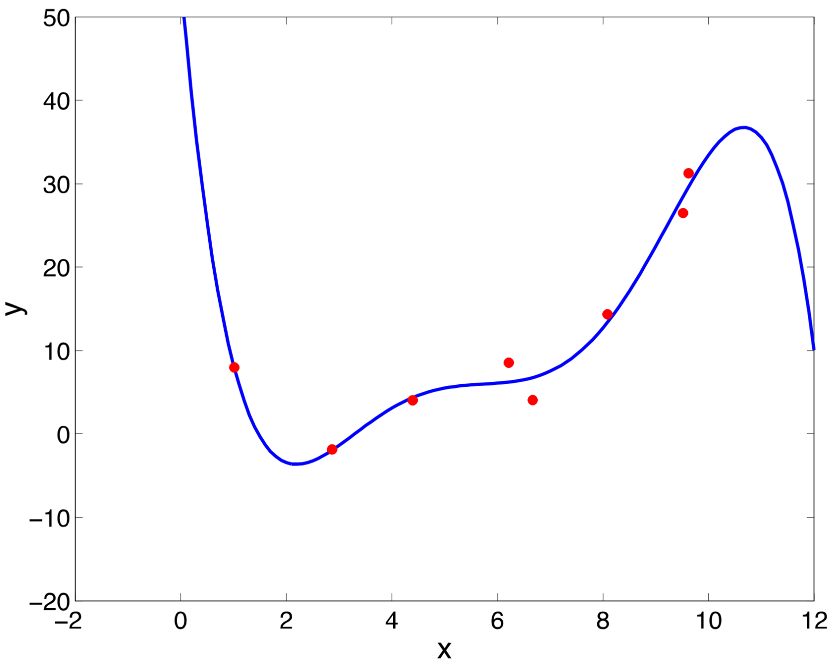
$$y^{(n)} = \omega_0 + \omega_1 x^{(n)} + \omega_2 x^{(n)} \ldots + \omega_2 x^{(n)} + \epsilon$$

where data noise is Gaussian distributed, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

We denote with θ the **unknown parameters**,

$$\theta = (\omega_0, \ldots, \omega_m, \sigma)$$

Unknown's



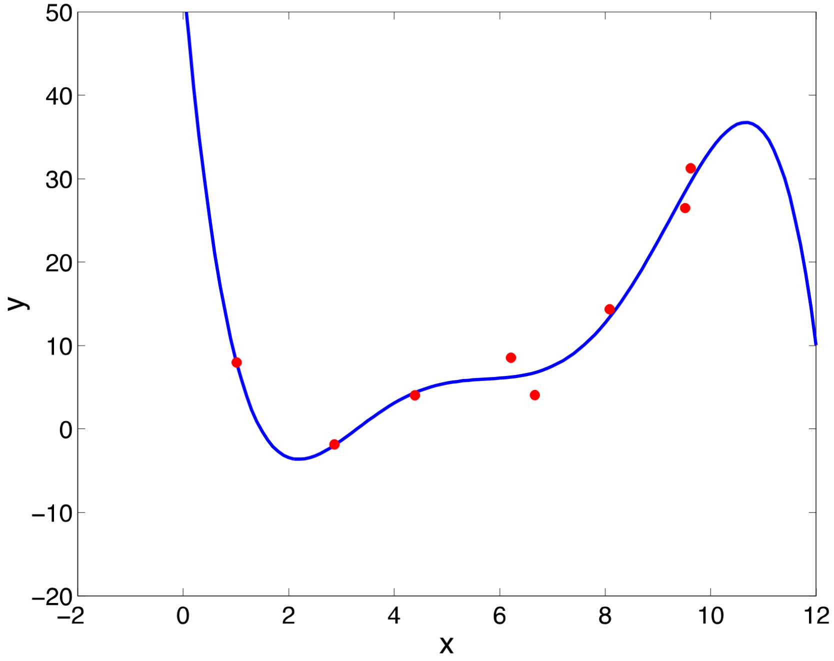
Bayesian modelling

Example of a linear model

Data: $\mathcal{D} = \{x^{(n)}, y^{(n)}\}$

Model: $y^{(n)} = \omega_0 + \omega_1 x^{(n)} + \omega_2 x^{(n)} \ldots + \omega_2 x^{(n)} + \epsilon$

Unknown parameters: $\theta = (\omega_0, \ldots, \omega_m, \sigma)$



Goal: To infer $\boldsymbol{\theta}$ from the data and to predict future outputs $p(y|x, \boldsymbol{\theta}, \boldsymbol{\mathcal{D}})$

Get θ ; $p(y|x, \theta, \mathcal{D})$

Bayesian modelling

$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$: likelihood of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta})$: prior probability of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$: posterior of $\boldsymbol{\theta}$ given $\boldsymbol{\mathcal{D}}$

$p(\boldsymbol{\mathcal{D}})$: marginal probability of $\boldsymbol{\mathcal{D}}$

$p(y|x, \boldsymbol{\mathcal{D}})$: predictive distribution

Bayesian modelling

[Prior knowledge] (Prior) Model dependent probability.

Marginal Data dependent Probability

How likely is Θ $p(D|\theta)$

$p(\mathcal{D}|\boldsymbol{\theta})$: likelihood of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta})$: prior probability of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta}|\mathcal{D})$: posterior of $\boldsymbol{\theta}$, given $\mathcal{D}$

$p(\mathcal{D})$: marginal probability of $\mathcal{D}$

$p(y|x, \mathcal{D})$: predictive distribution

Bayes' rule:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{P(\mathcal{D})}$$

Posterior (Model Given Data)

Likelihood (Data Given Model)

Prediction of a new point:

$$p(y|x, \mathcal{D}) = \int p(y|\theta, x, \mathcal{D}) p(\theta|\mathcal{D})\, d\theta$$

Rewrite Quantities.

$$PP = \frac{L \cdot P}{MP}$$

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)}.$$

Prediction: $p(y|x, D) = \int p(y|\theta, x, D)\, p(\theta|D)\, d\theta$

"Prediction using posterior" [weighted posterior probabilities]

Bayesian modelling

Observations

- In contrast to the maximum likelihood estimation (MLE), in Bayesian learning **we average over possible parameter settings** rather than optimizing over parameter space.
- Bayesian inference gives us a **systematic way to express our uncertainty** about future predictions. Prediction is not just a point estimate (as for MLE) but has a probability form that expresses the uncertainty about the predictions.

Express Uncertainity?
Probablistic approach.

MLE → optimizing parameter space
(Point estimate)

Bayesian → Average over parameters.
(Probability)

Bayesian Inference

Bayesian Model Selection

Commonalities

Bayesian

Occam

The principle of Occam's Razor

The **principle of Occam's razor** in its original formulation states that:

"Entia non sunt multiplicanda praeter necessitatem"

(In English, "Entities should not be multiplied unnecessarily")

Many scientists have adopted or reformulated the Occam's Razor principle, which is often cited in stronger forms, as in the following statement:

- "If you have two theories that both explain the observed facts, then you should use the simplest until more evidence comes along"
- "One should pick the simplest model that adequately explains the data"

Both should explain

Picture from: https://www.britannica.com/topic/Occams-razor

The model Selection problem

different batch for each iteration.

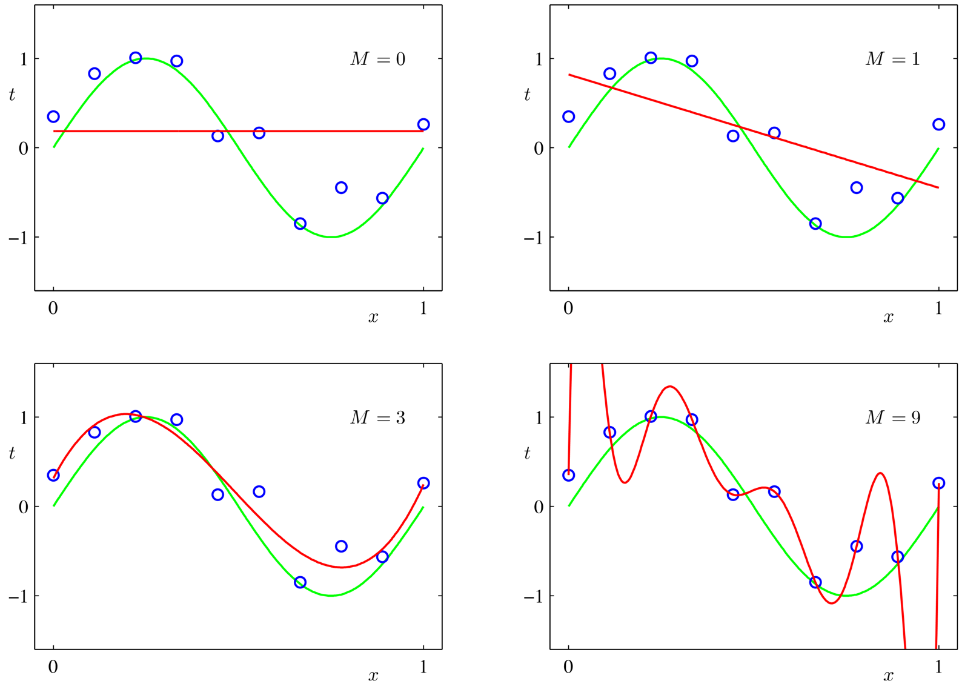
We could perform K-fold cross-validation (CV) to estimate the generalization error of all candidates model.

However, it requires fitting each candidate model K times!

→ A more efficient approach is given by Bayesian modelling

?

Polynomial Regression.



Which of the above models represents data the best?

Occam's razor and Bayesian model selection

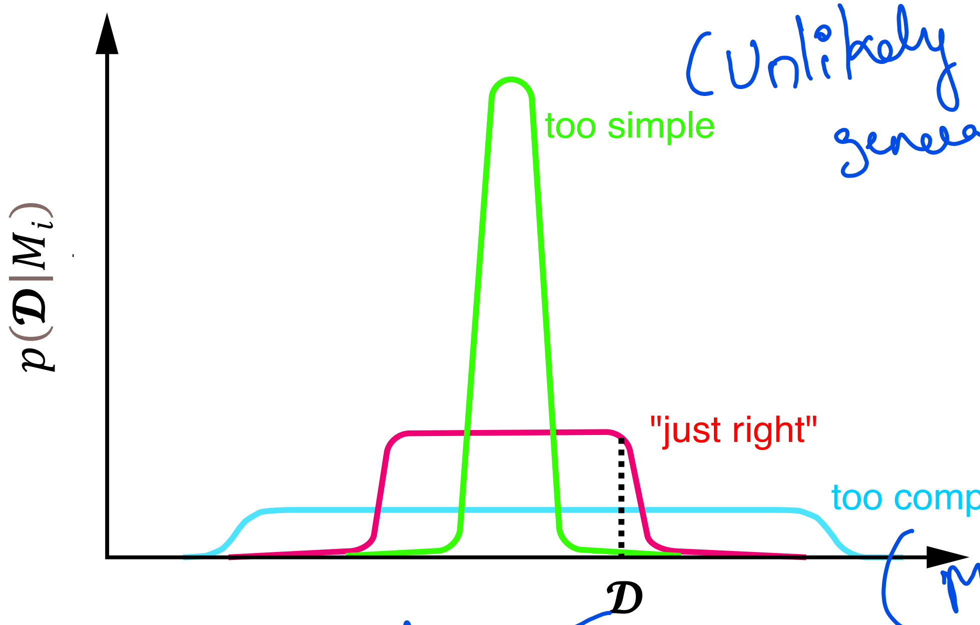
Any model. Comparing Models;

Marginal ⇒ Denominator.

We can compare different models using the marginal likelihood:

$$p(\boldsymbol{\mathcal{D}}|M_i) = \int p(\boldsymbol{\mathcal{D}}|\theta, M_i)p(\theta|M_i)\, d\theta$$

- Model classes that are **too simple** are unlikely to generate the data set.
- Model classes that are **too complex** can generate many possible data sets, so again, they are unlikely to generate that particular data set $\boldsymbol{\mathcal{D}}$.



(Unlikely to generate data)

(Many data sets)

Actual Dataset.

(Highest for just right)

Image from: Rasmussen, C., & Ghahramani, Z. (2000). Occam's razor. Advances in neural information processing systems, 13.

Occam's razor and Bayesian model selection

To understand the Bayesian Occam's razor, we notice that:

$$\int_{\boldsymbol{\mathcal{D}}} p(\boldsymbol{\mathcal{D}}|M_i) = 1$$

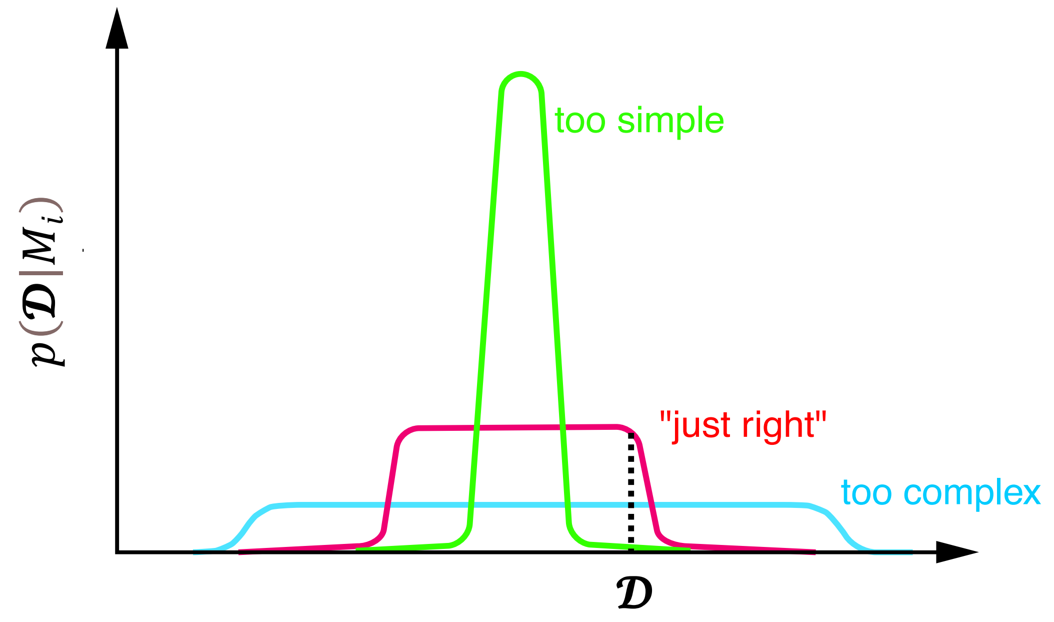


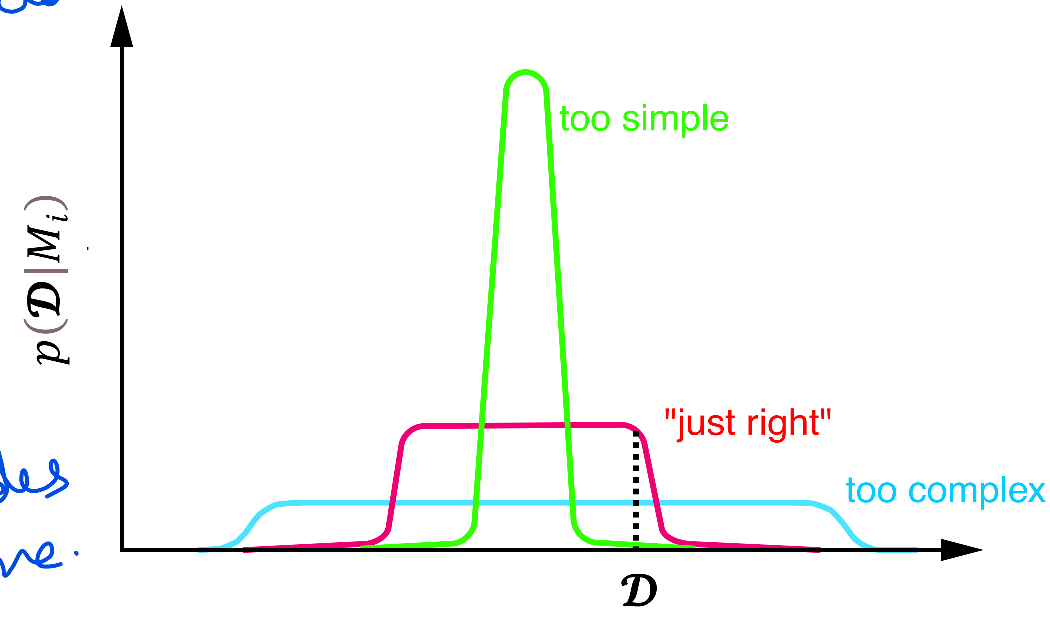
Image from: Rasmussen, C., & Ghahramani, Z. (2000). Occam's razor. Advances in neural information processing systems, 13.

Occam's razor and Bayesian model selection

To understand the Bayesian Occam's razor, we notice that:

Marginal probabilities over dataset sum up to 1.

$$\int_{\mathcal{D}} p(\mathcal{D}|M_i) = 1$$



Intuitively, complex models which can predict *many* datasets, must spread their probability mass
→ They don't attribute large probability for any given data set as simpler models.

Jack of All trades Master of None.

Image from: Rasmussen, C., & Ghahramani, Z. (2000). Occam's razor. Advances in neural information processing systems, 13.

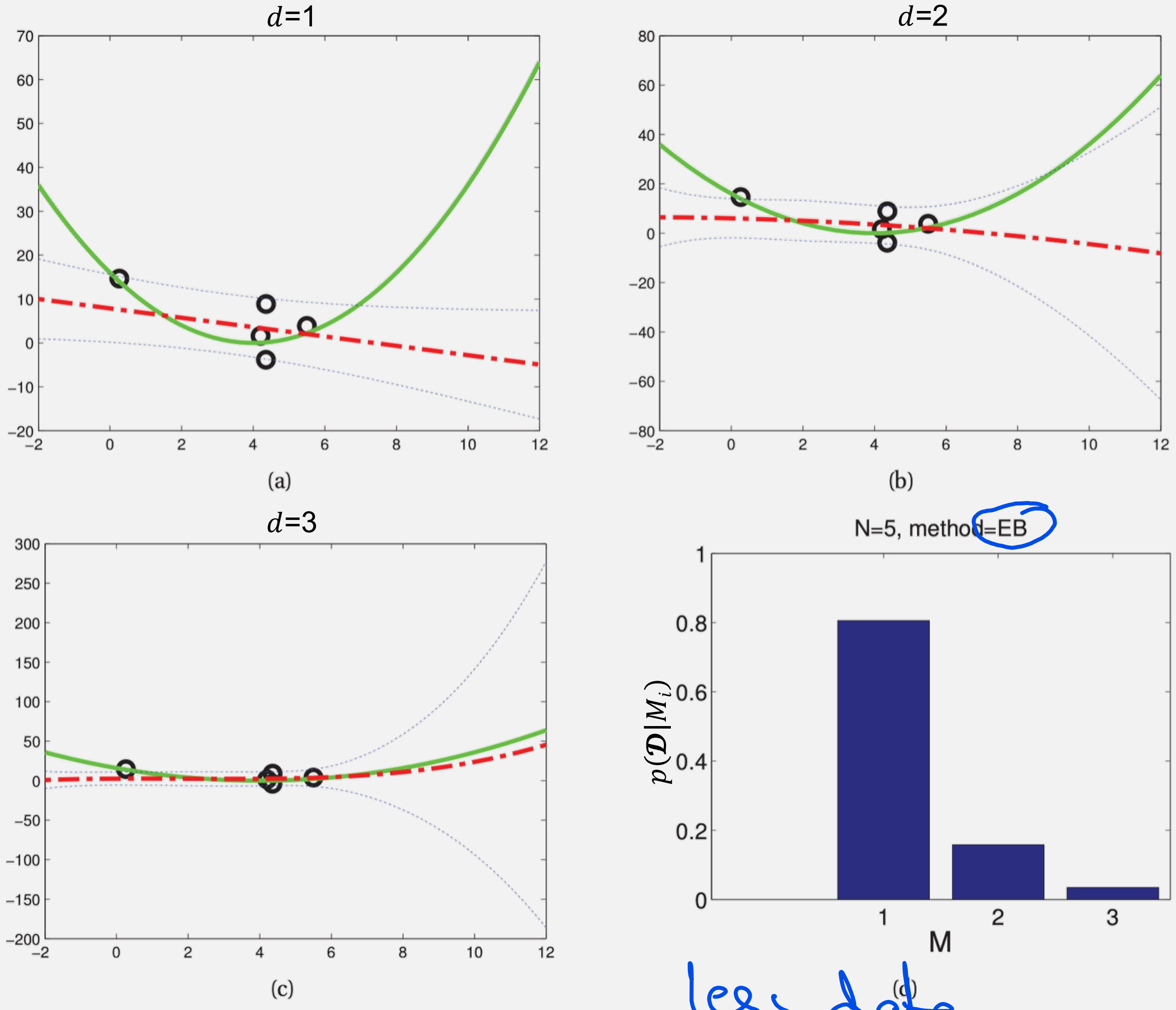
Bayesian Occam's razor

A concrete example

We plot polynomials of degrees 1, 2 and 3 fit to N=5 data points using (empirical) Bayes.

- True function
- Prediction
- $\pm\sigma$ around the mean

There is not enough data to justify a complex model, so the best model is d = 1.



less data.

Bayesian Occam's razor

A concrete example

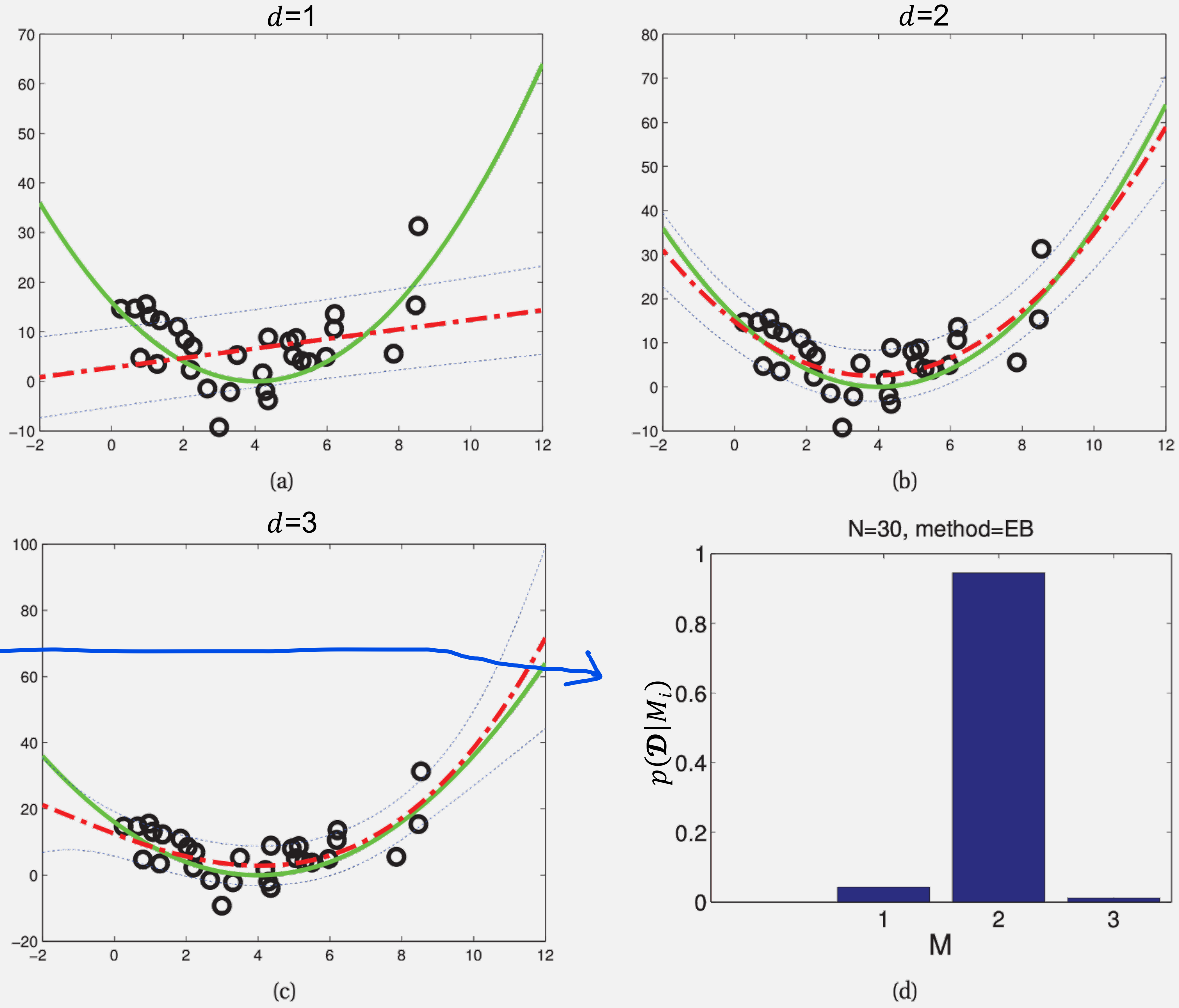
We plot polynomials of degrees 1, 2 and 3 fit to N=30 data points using (empirical) Bayes.

More sample points.

- True function
- Prediction
- $\pm\sigma$ around the mean

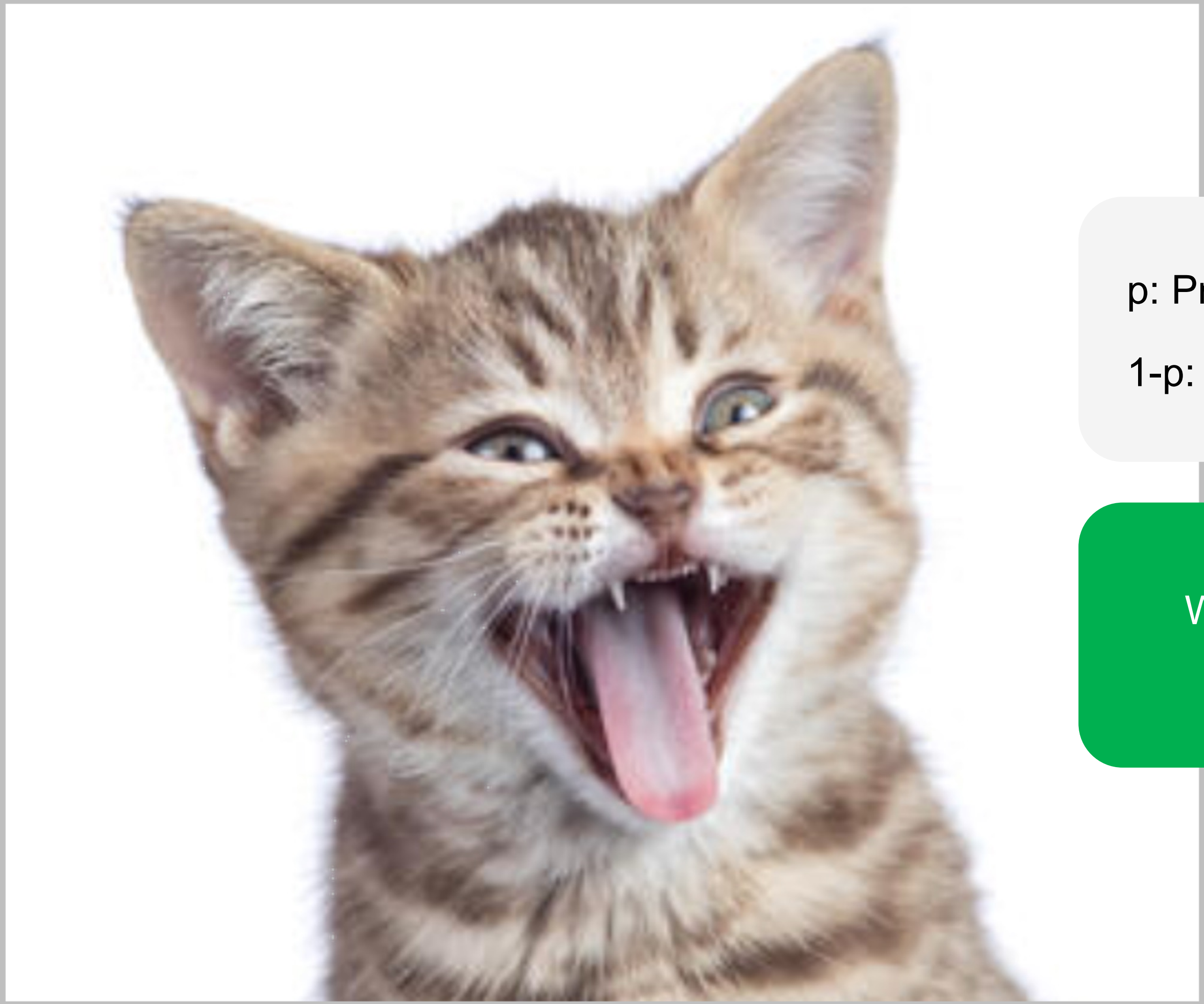
reflected here

When more data is available, d = 2 is the right model



Bayesian Inference

Prior Distributions



p: Prob. purring

1-p: Prob. grumpy

What is the best guess for the probability p?

How can I update my belief on p?

Prior distributions

The importance of priors in Bayesian Inference

$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$: likelihood of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta})$: prior probability of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$: posterior of $\boldsymbol{\theta}$, given $\boldsymbol{\mathcal{D}}$

$p(\boldsymbol{\mathcal{D}})$: marginal probability of $\boldsymbol{\mathcal{D}}$

$p(y|x,\boldsymbol{\mathcal{D}})$: predictive distribution

Bayes' rule:

$$p(\theta|\boldsymbol{\mathcal{D}}) = \frac{p(\boldsymbol{\mathcal{D}}|\theta)\ p(\theta)}{P(\boldsymbol{\mathcal{D}})}$$

Prior distributions

The importance of priors in Bayesian Inference

Prior beliefs ?

$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$: likelihood of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta})$: prior probability of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$: posterior of $\boldsymbol{\theta}$, given $\boldsymbol{\mathcal{D}}$

$p(\boldsymbol{\mathcal{D}})$: marginal probability of $\boldsymbol{\mathcal{D}}$

$p(y|x,\boldsymbol{\mathcal{D}})$: predictive distribution

Prior probability

Bayes' rule:

$$p(\theta|\boldsymbol{\mathcal{D}}) = \frac{p(\boldsymbol{\mathcal{D}}|\theta)\ \boxed{p(\theta)}}{P(\boldsymbol{\mathcal{D}})}$$

→ before evidence is taken

A **prior probability distribution** of an uncertain quantity is the probability distribution that would express one's belief, before some evidence is taken into account.

→ For example, a prior could represent the distribution of votes coming from an opinion poll, prior to the election.

Priors: Subjective vs. Objective

Machine Learning
Data Analytics

Assumptions.

A **subjective prior** expresses the modeler's subjective belief.

- We formulate our (subjective) assumptions about modeling the data in terms of priors
- We have to work hard to understand the system under study in order to formulate our assumptions

→ uninformative

An **objective prior** constrain prior beliefs to be "uninformative" about the parameters.

- The objective Bayes view is that formulating our assumptions is too difficult, especially in complex models

↳ Very difficult

Let data speak for itself.

If we don't have strong beliefs about what θ should be, it is common to use an "uninformative" priors → **"Let the data speak for itself!"**

Priors: Informative vs. Uninformative

Specific

An **informative prior** expresses a specific information about a variable.

- For example, a reasonable informative prior about the temperature at noon tomorrow could be given by a normal distribution with expeced value equal to today's noon temperature and variance equal to the daily variance of the temperature.

vague or general.

An **uninformative prior** is designed to express vague or general information about a variable.

- For example, when tossing a coin, we assign the probability of 0.5 to both heads and tails.

→ Vague.

No information

Some inform

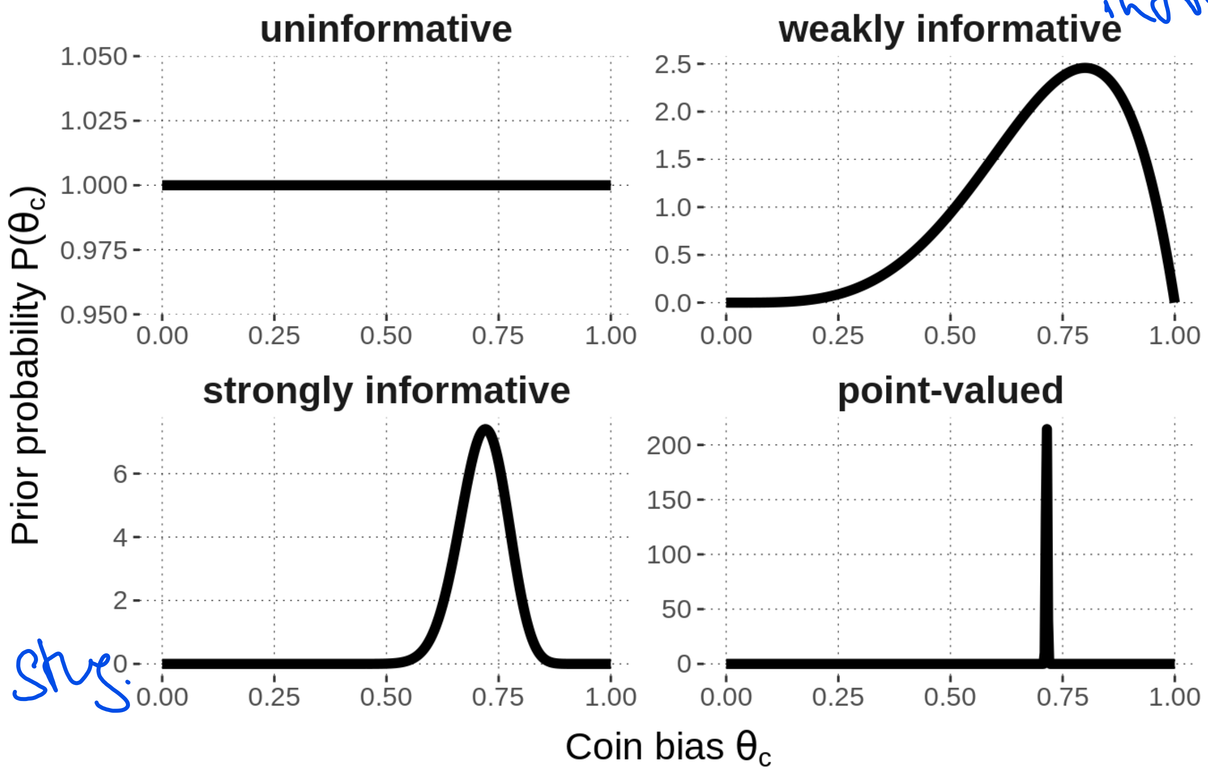


Image from: https://michael-franke.github.io/intro-data-analysis/Chap-03-03-models-parameters-priors.html

Hierarchical priors

The multi-level model

A key requirement for computing the posterior $p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$ is the specification of a prior $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$, where alpha denotes a set of hyperparameters.

Hyperparameters → Prior $\theta | \alpha$

posterior $p(\theta|\mathcal{D})$

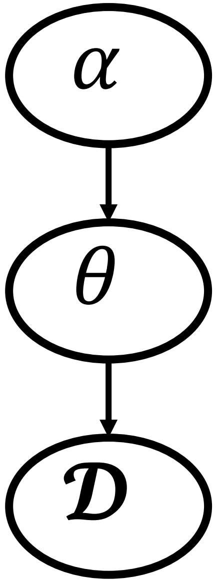
We have multiple levels of priors:

$$\alpha \rightarrow \theta \ \rightarrow \boldsymbol{\mathcal{D}}$$

$$p(\theta) = \int p(\theta|\alpha)\, p(\alpha)\, d\alpha$$

$$p(\boldsymbol{\mathcal{D}}) = \int p(\theta)\, p(\boldsymbol{\mathcal{D}}|\theta)\, \mathrm{d}\theta$$

The multi-level model



Hierarchical priors

An example: modeling cancer mortality in various cities

Consider the problem of predicting the cancer mortality rates in various cities. We measure the number of people N_i in various cities, as well as the number of people who died of cancer x_i in those cities. We assume that the mortality follows:

$$x_i \sim Bin(N_i, \theta_i).$$

A reasonable approach to estimate θ_i is that of assuming that they are drawn from the same distribution $\theta_i \sim Beta(a, b)$, where $\alpha = (a, b)$ are hyper-parameters in our model.

Then, the full joint distribution is written as

$$p(\boldsymbol{\mathcal{D}}, \theta, \alpha | N) = p(\alpha) \prod_{i=1}^{N} Bin(x_i | N_i, \theta_i) Beta(\theta_i | \alpha).$$

Note: It is crucial to infer α from the data itself. ← Hyper parameter from data itself

Example taken from: Machine Leraning: A Probabilistic Perspective, Ch. 5.51

Empirical Prior

A computational "shortcut"

In hierarchical models, we need to **compute the posterior on multiple levels of variables**. For example,

$$p(\alpha, \theta|\boldsymbol{\mathcal{D}}) \propto p(\boldsymbol{\mathcal{D}}|\theta)\, p(\theta|\alpha)\, p(\alpha).$$

In some case, we can simplify the problems by **marginalizing over $\boldsymbol{\theta}$**. Then, we just need to compute:

$$p(\alpha|D).$$

approximate posterior with a point estimate

As a computational "shortcut", we can approximate the posterior on the hyper-parameters α with a point estimate. Since α is generally of a much smaller dimensionality than θ, it is less prone to overfitting and we safely assume a uniform prior on α.

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, p(\boldsymbol{\mathcal{D}}, \alpha) = \underset{\alpha}{\operatorname{argmax}} \int p(\boldsymbol{\mathcal{D}}, \theta) p(\theta|\alpha) d\theta$$

Conjugate Prior

A computational "shortcut"

same algebraic form.

likelihood.
p(y|θ)

A prior $p(\theta)$ is a **conjugate prior** for a particular likelihood $p(y|\theta)$ if the resulting posterior $p(\theta|y)$ has the same algebraic form.

Conjugate priors are widely used because they provide advantages:

- they usually allow us to derive a closed-form expression for the posterior distribution;
- they are easy to interpret,

Not flexible

Note: Conjugate priors simplify the computation, but are often not flexible enough to encode our prior knowledge → We can also use mixture of conjugate priors.

mixture

Conjugate Prior

An example

Likelihood (Binomial): $p(\mathcal{D}|\boldsymbol{\theta}) = Bin(\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$

Prior (Beta): $p(\boldsymbol{\theta}) = Beta(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

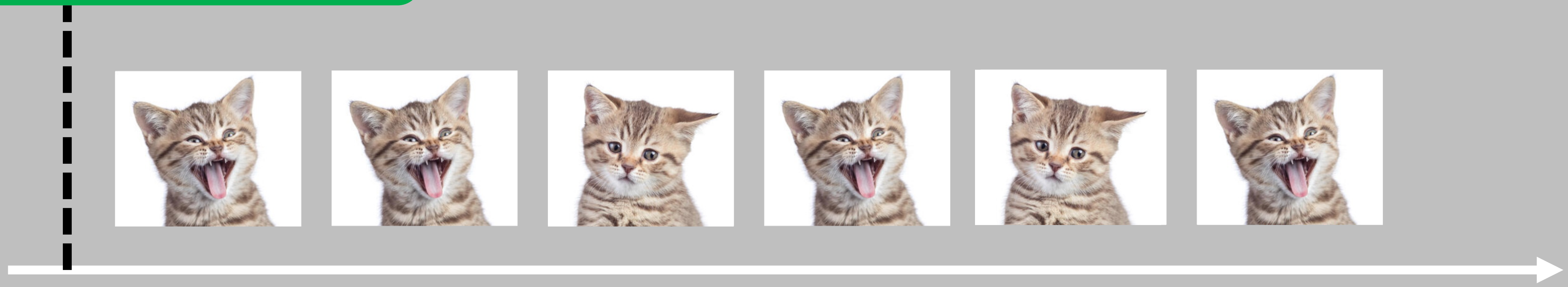
We plug them into the Bayes' formula to derive the posterior distribution:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\; p(\theta)}{\int p(\mathcal{D}|\theta)\;\; p(\theta) d\theta}$$

$$= \frac{\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\;\; \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \binom{n}{x}\theta^{x}(1-\theta)^{n-x}\;\; \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\; d\theta}$$

$$= \frac{\frac{n\,C_x}{B(\alpha,\beta)}\;\theta^{x+\alpha-1}\,(1-\theta)^{n-x+\beta-1}}{\frac{n\,C_x}{B(\alpha,\beta)}\;\int \theta^{x+\alpha-1}\,(1-\theta)^{n-x+\beta-1}\;\; d\theta} = Beta(x+\alpha, n-x+\beta)$$

Prior: Beta(2, 2)



Prior: Beta(2, 2)

Posterior: Beta(2+2, 4+2) = Beta(4, 6)

Prior: Beta(2, 2)

Posterior: Beta(2+2, 4+2) = Beta(4, 6)

Example inspired by: https://towardsdatascience.com/understanding-conjugate-priors-21b2824cddae

Bayesian Inference

Linear Regression (Bayesian treatment)

Linear Regression (Bayesian treatment)

Machine Learning
Data Analytics

Given the observed data $\mathcal{D} = \{x^{(n)}, y^{(n)}\}$, we assume to know the noise variance σ^2.

We would like to compute the posterior over the parameters, i.e,

$$p(w|\mathcal{D}, \sigma^2).$$

(We assume throughout a Gaussian likelihood model).

In linear regression **the likelihood is given by:**

$$p(y|X, w, \mu, \sigma^2) = \mathcal{N}(y|\mu + Xw,\ \sigma^2 I_N)$$

$$\propto \exp(-\frac{1}{2\sigma^2}(y - \mu - Xw)^T(y - \mu - Xw))$$

where μ is an offset term.

$\hookrightarrow$ bias / offset term.

Posterior over parameters.

$\mu = \mu + Xw.$

$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Linear Regression (Bayesian treatment)

Machine Learning Data Analytics FAU

The conjugate prior of a Gaussian likelihood is also Gaussian*, which we will denote by

$$p(w) = \mathcal{N}(w|w_0, V_0).$$

Using the Bayes rule for Gaussian*, the posterior is given by

$$p(w|X, y, \sigma^2) \;\propto \mathcal{N}(w|w_0, V_0)\,\mathcal{N}(\mathrm{y}|\mathrm{Xw}, \sigma^2 I_N) = \mathcal{N}(w|w_N, V_N)$$

where

$$w_N = \; V_N V_0^{-1} w_0 + \frac{1}{\sigma^2} V_N X^T y$$

$$V_N = \; \sigma^2(\sigma^2 V_0^{-1} + \; X^T X)^{-1}$$

variance in parameters.

* See: Murphy K., „Machine Learning: A Probabilistic Perspective“ (2012)

Linear Regression (Bayesian treatment)

The posterior predictive distribution at a test point x is given by *

$$p(y|x, \boldsymbol{\mathcal{D}}, \sigma^2) = \int \mathcal{N}(y|x^T w, \sigma^2)\mathcal{N}(w|w_N, V_N)dw$$

$$= \mathcal{N}(y|w_N^T x, \sigma_N^2(x))$$

where $\sigma_N^2(x) = \sigma^2 + x^T V_N x$.

The variance in this prediction depends on the variance of the observation noise, σ^2, and the variance in the parameters, V_N.

Bayesian Inference

Recap

Recap

- Bayesian modelling
  - Prior
  - Posterior
  - Likelihood
- Bayesian modelling requires integration over parameters
  - For complex models it could be not tractable! (We cannot compute the integral analytically)
- Priors
  - Informative vs Uninformative
  - Conjugate priors
- Linear regression with Bayesian treatment