UNIVERSITY OF
OXFORD

Lecture 2: Bayesian Modeling (Part 1)

Advanced Topics in Machine Learning

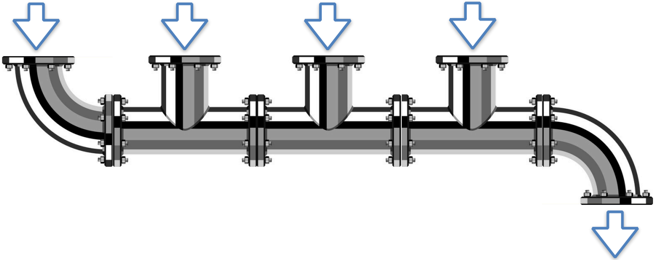
Dr. Tom Rainforth

January 24nd, 2020

rainforth@stats.ox.ac.uk

Last Lecture Recap: The Bayesian Pipeline

$\overbrace{p(\theta)}^{\text{Prior}}$ $\overbrace{p(\mathcal{D}|\theta)}^{\text{Likelihood}}$ $\overbrace{\mathcal{D}}^{\text{Data}}$ Inference Method



$$\underbrace{p(\theta|\mathcal{D})}_{\text{Posterior}} \propto p(\mathcal{D}|\theta)p(\theta)$$

Lecture 2 Outline

- What is a Bayesian model?
- Bayesian modeling through the eyes of multiple hypotheses
- Worked example: Bayesian linear regression
- What makes a good model and how do we compare between models?
- Bayesian model averaging

What is a Bayesian Model?

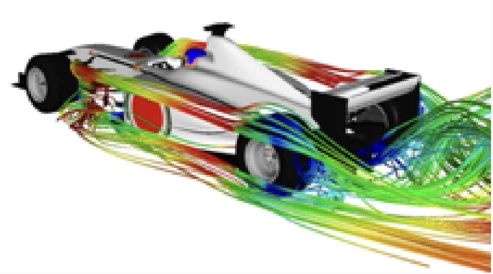
What is a Model?

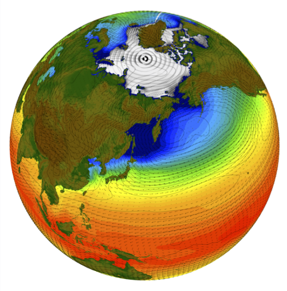
- Models are mechanisms for **reasoning** about the world
- Examples: Newtonian mechanics, simulators, internal models our brain constructs
- Good models balance **fidelity**, **predictive power** and **tractability**
    - Quantum mechanics is a more accurate model than Newtonian mechanics, but it is actually less useful for everyday tasks

Example Model: Poker Players Reasoning about Each Other

©DANNYMAXWELL
Bad cards, should probably fold
Can I bluff him though? I doubt he has good cards
Cards are ok, but if he bets I'll fold
He might be bluffing though...

Example Model: Computer Simulations

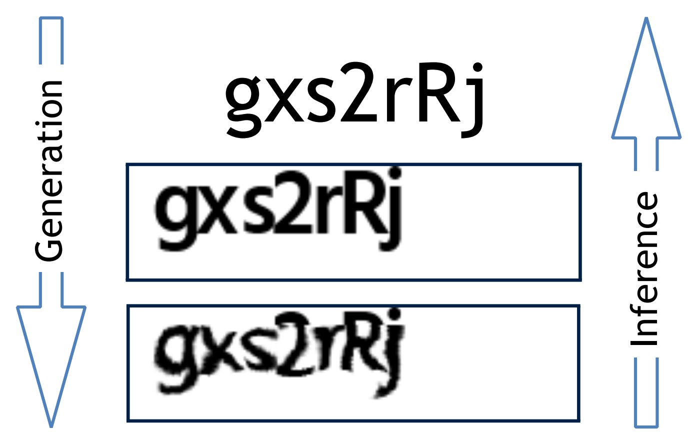




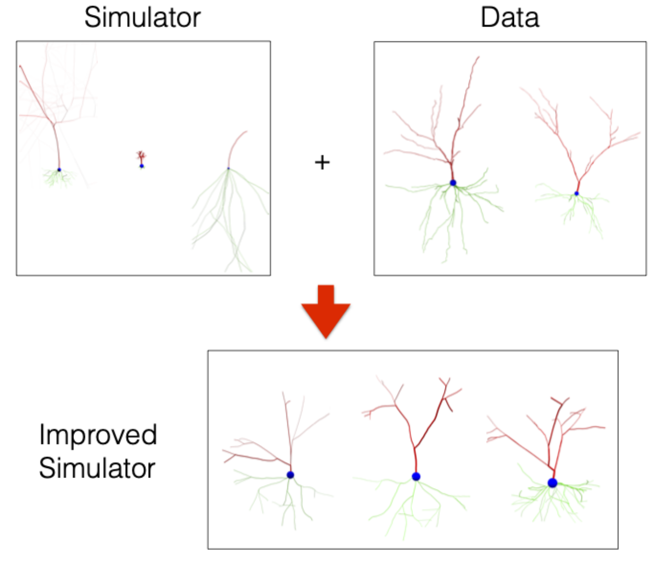
What is Bayesian Model?

- A Bayesian model is a **probabilistic generative model** $p(\theta, \mathcal{D})$ over **latents** θ and **data** $\mathcal{D}$
- It forms a probabilistic "simulator" for generating data that we **might** have seen
- Pretty much any stochastic simulator can be used as a Bayesian model (we will return to this idea in more detail when we cover probabilistic programming)

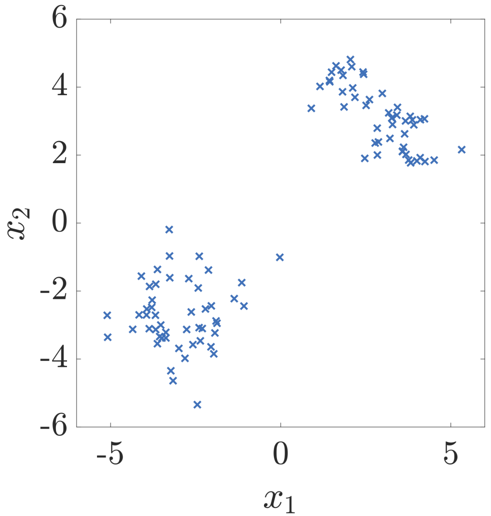
Example Bayesian Model: Captcha Simulator



Example Bayesian Model: Neuron Growth



Example Bayesian Model: Gaussian Mixture Model



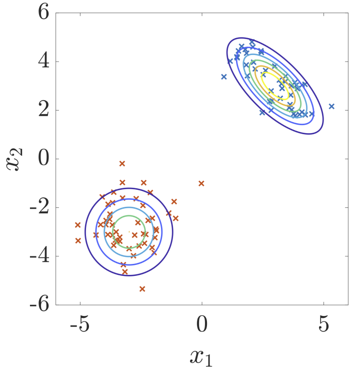
Example Bayesian Model: Gaussian Mixture Model

Fixed parameters:

$$\pi = [0.5, 0.5]$$

$$\mu_1 = [-3, -3] \quad \mu_2 = [3, 3]$$

$$\Sigma_1 = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}, \quad \Sigma_2 \quad = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Generative model:

$$\theta \sim \text{Categorical}(\pi)$$

$$x \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

Generative model (full dataset):

$$p(\theta, \mathcal{D}) = \prod_{n=1}^{N} p(\theta_n, x_n)$$

A Fundamental Assumption

- An assumption made by virtually all Bayesian models is that datapoints are conditionally independent given the parameter values.
- In other words, if our data is given by $\mathcal{D} = \{x_n\}_{n=1}^{N}$, we assume that the likelihood factorizes as follows

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(x_n|\theta). \tag{1}$$

- This effectively equates to assuming that our model captures all information relevant to prediction
- For more details, see the notes

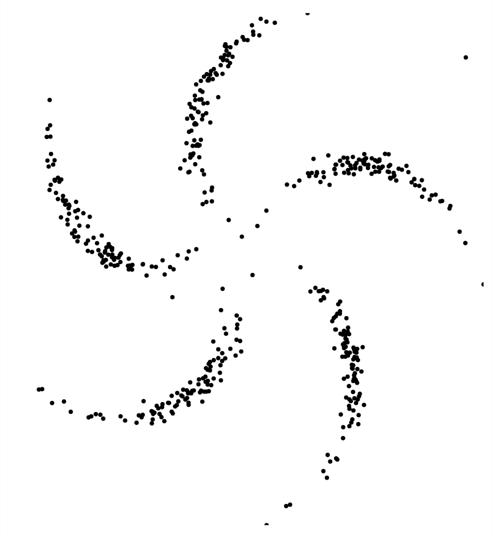
All models are wrong, but some are useful

—George Box

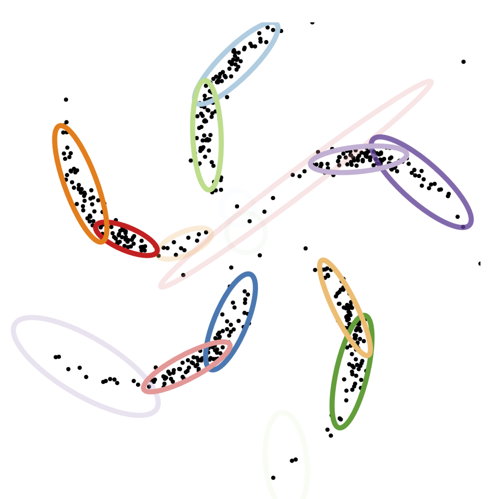
What is the Purpose of a Model?

- The purpose of a model is to help provide insights into a target problem or data and sometimes to further use these insights to make predictions
- Its purpose is **not** to try and fully encapsulate the "true" generative process or perfectly describe the data
- There are infinite different ways to generate any given dataset
    - Trying to uncover the "true" generative process is not even a well-defined problem
- In any real–world scenario, no Bayesian model can be "correct"
    - The posterior is inherently subjective
- It is still important to criticize—models can be very wrong!
    - E.g. we can use frequentist methods to falsify the likelihood

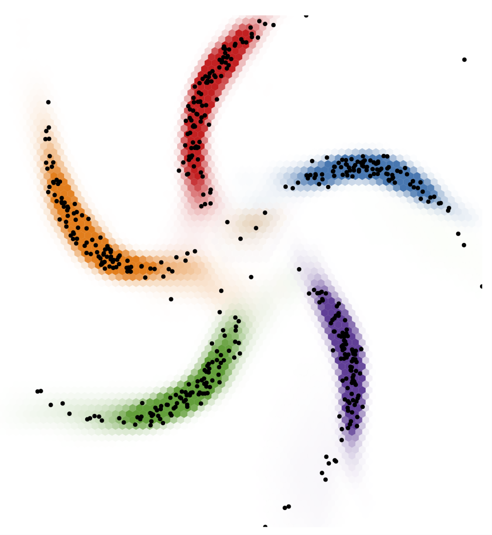
Some Models are Much Better than Others



Some Models are Much Better than Others



Some Models are Much Better than Others



Bayesian Modeling Through the Eyes of Multiple Hypotheses

Bayesian Modeling as Multiple Hypotheses

Bayesian models are rooted in **hypotheses**:

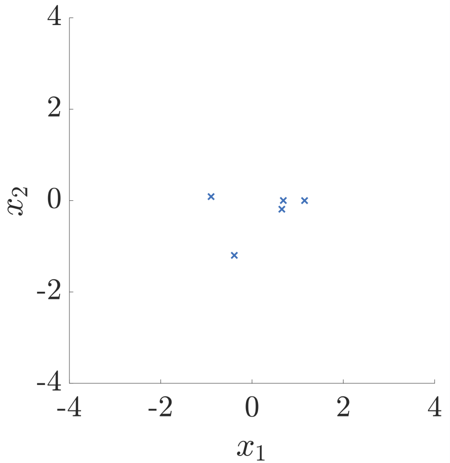
- Each instance of our parameters θ is a hypothesis. Given a θ, we can simulate data using the likelihood model $p(\mathcal{D}|\theta)$
- **Bayesian inference** allows us to reason about these hypothesis, giving the probability that each is true given the actual data we observe
- The posterior predictive is a weighted sum of the predictions from all possible hypotheses, where these weights are how likely that hypothesis is to be true

Example: Density Estimation

Presume that we decide to use an isotropic Gaussian likelihood with unknown mean θ to model the data on the right:

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} \mathcal{N}(x_n; \theta, I)$$

where I is a two-dimensional identity matrix

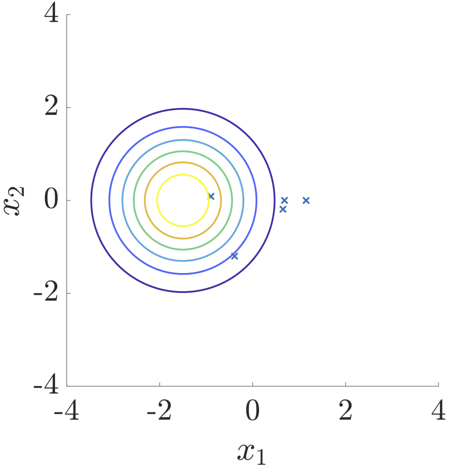


Example: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

$p(\mathcal{D}|\theta = [-2, 0])$

$= 0.00059 \times 10^{-5}$



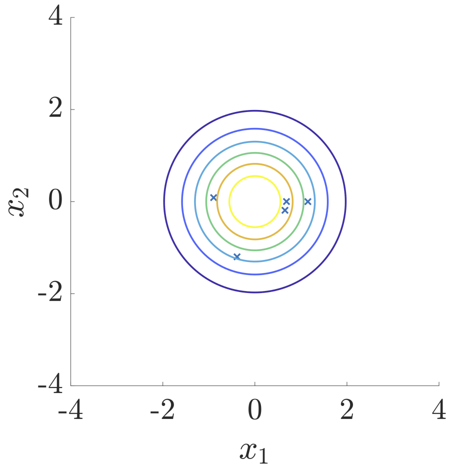
Example: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

$$p(\mathcal{D}|\theta = [-2, 0]) \\ = 0.00059 \times 10^{-5}$$

Hypothesis 2: $\theta = [0, 0]$

$$p(\mathcal{D}|\theta = [0, 0]) \\ = 0.99 \times 10^{-5}$$



Example: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

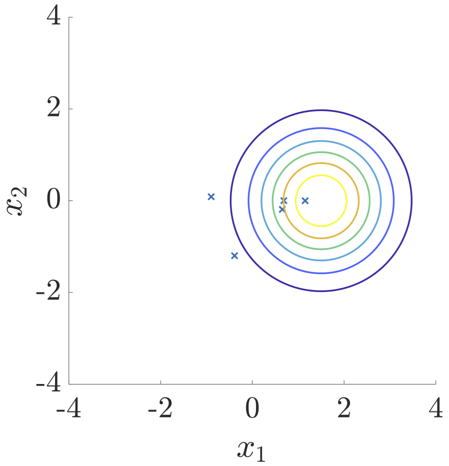
$$p(\mathcal{D}|\theta = [-2, 0])$$
$$= 0.00059 \times 10^{-5}$$

Hypothesis 2: $\theta = [0, 0]$

$$p(\mathcal{D}|\theta = [0, 0])$$
$$= 0.99 \times 10^{-5}$$

Hypothesis 3: $\theta = [2, 0]$

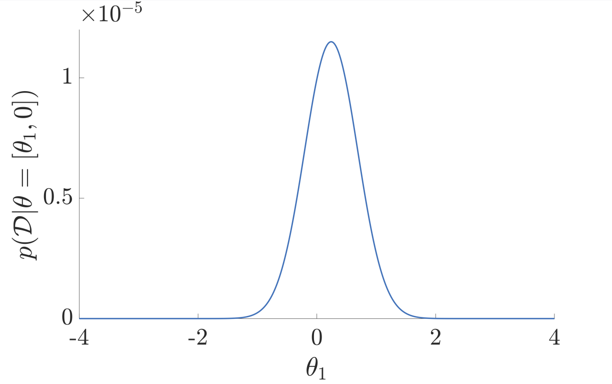
$$p(\mathcal{D}|\theta = [2, 0])$$
$$= 0.021 \times 10^{-5}$$



Example: Density Estimation (2)

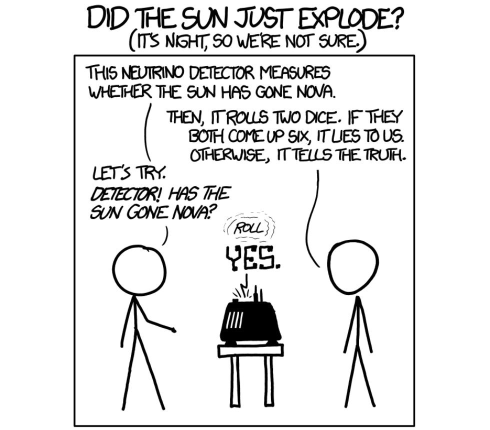
Clearly $\theta = [0, 0]$ is the best of these three hypotheses.

More generally, the likelihood model is telling us how **likely** each hypothesis is to be correct given the data we observe.



Bayesian Modeling Allows us To Express our Prior Beliefs

- In the above example we only considered the likelihood of each hypothesis
- We may though have unequal prior beliefs about each hypothesis


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.

https://xkcd.com/1132/

The Posterior Predictive Averages over Hypotheses

The posterior predictive distribution allows us to **average** over each of our hypotheses, weighting each by their posterior probability.

For example, in our density estimation example, lets introduce (the rather unusual but demonstrative) prior,

$$p(\theta) = \begin{cases} 0.05 \quad \text{if} \quad \theta = [-2, 0] \\ 0.05 \quad \text{if} \quad \theta = [0, 0] \\ 0.9 \quad \text{if} \quad \theta = [2, 0] \\ 0 \quad \text{otherwise} \end{cases} \tag{2}$$

The Posterior Predictive Averages over Hypotheses (2)

Then we have (note $d\theta$ is a counting measure below)

$$\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\theta)p(\theta|\mathcal{D})d\theta \\
&= \frac{1}{p(\mathcal{D})}\int p(x|\theta)p(\theta,\mathcal{D})d\theta \\
&= \frac{1}{p(\mathcal{D})}\Big(\mathcal{N}(x;[-2,0],I)\times 0.05\times p(\mathcal{D}|\theta=[-2,0]) \\
&\qquad + \mathcal{N}(x;[0,0],I)\times 0.05\times p(\mathcal{D}|\theta=[0,0]) \\
&\qquad + \mathcal{N}(x;[2,0],I)\times 0.9\times p(\mathcal{D}|\theta=[2,0])\Big)
\end{aligned}$$

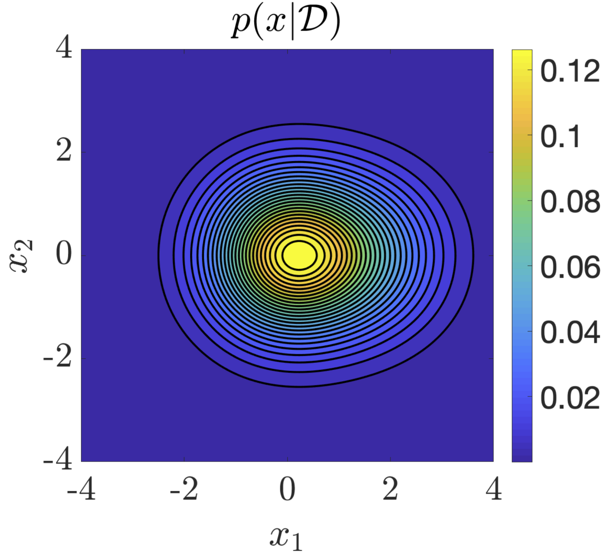
The Posterior Predictive Averages over Hypotheses (3)

Inserting our likelihoods from earlier and trawling through the algebra now gives

$$\begin{aligned} p(x|\mathcal{D}) = & 0.0004 \times \mathcal{N}(x; [-2, 0], I) \\ & + 0.716 \times \mathcal{N}(x; [0, 0], I) \\ & + 0.283 \times \mathcal{N}(x; [2, 0], I) \end{aligned}$$

We thus have that the posterior predictive is a weighted sum of the three possible predictive distributions

The Posterior Predictive Averages over Hypotheses (4)



An Important Subtlety

- Even though we average over θ, a Bayesian model is still implicitly assuming that there is still a single true θ
    - The averaging over hypotheses is from our own uncertainty as the which one is correct
    - This can be problematic with lots of data given our model is an approximation
- In the limit of large data, the posterior is guaranteed to collapse to a point estimate:

$$p(\theta|x_{1:N}) \rightarrow \delta(\theta = \hat{\theta}) \quad \text{as} \quad N \rightarrow \infty \tag{3}$$

- The value of $\hat{\theta}$ and the exact nature of this convergence is dictated by the Bernstein–von Mises Theorem (see the notes)
- Note that, subject to mild assumptions, $\hat{\theta}$ is independent of the prior
    - With enough data, the likelihood always dominates the prior

Worked Example: Bayesian Linear Regression

Linear Regression

House size is a good linear predictor for price (ignore the colors)

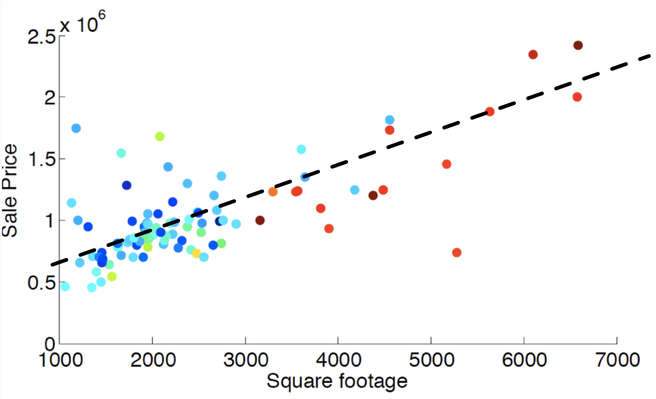


Image credit: Pier Palamara

Linear Regression (2)

Here we have:

- Inputs $x \in \mathbb{R}^D$ (where $D = 1$ for this particular problem)
- Outputs $y \in \mathbb{R}$
- Data $\mathcal{D}$ comprising of N input–output pairs: $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}$
- A regression model $y \approx x^T w + b$ where $w \in \mathbb{R}^D$ and $b \in \mathbb{R}$
- We can simplify this notation by redefining $x \leftarrow [1, x^T]^T$ and $w \leftarrow [b, w^T]^T$, such that we now have $y \approx x^T w$

Classical least squares linear regression is a discriminative method where we aim to minimize the empirical mean squared error

$$R = \frac{1}{N} \sum_{n=1}^{N} (y_n - x_n^T w)^2.$$

Linear Regression (3)

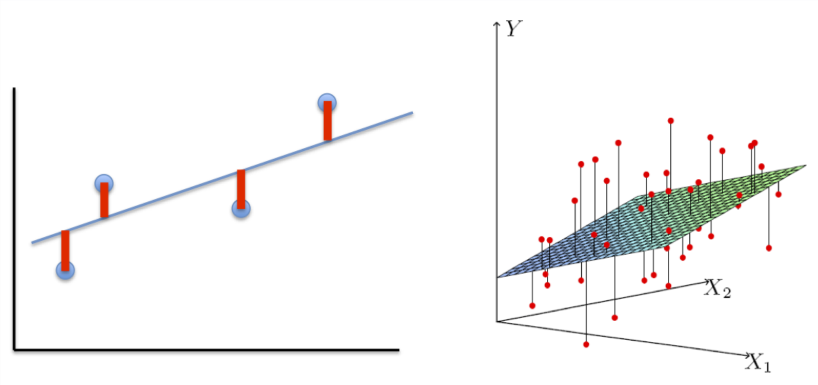


Image credit: Pier Palamara

Bayesian Linear Regression

- Defining $\mathbf{x} = [x_1, \ldots, x_N]^T$ and $\mathbf{y} = [y_1, \ldots, y_N]^T$, the least square solution is analytically given by (see last term's machine learning module for a derivation)

$$w^* = \left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T\mathbf{y} \tag{4}$$

- This only provides a point estimate for w
    - We have no **uncertainty** estimate
- We can introduce uncertainty by building a probabilistic generic model based around linear regression and then being Bayesian about the weights

Bayesian Linear Regression: Prior

The first step to do this is to define a prior over the weights. We will use a zero-mean Gaussian with a fixed covariance matrix C:

$$p(w) = \mathcal{N}(w; 0, C) \tag{5}$$

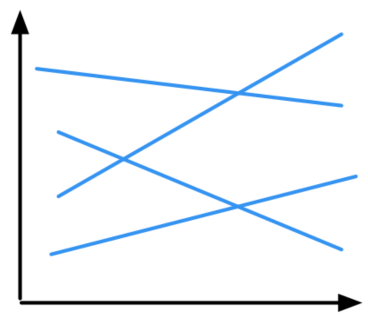


Image credit: Roger Grosse

Bayesian Linear Regression: Likelihood

We next need to introduce a likelihood model based on these weights. We will make the standard assumption that the datapoints are independent of each other given the weights and again use a Gaussian to give

$$p(\mathbf{y}|\mathbf{x}, w) = \prod_{n=1}^{N} p(y_n|x_n, w) = \prod_{n=1}^{N} \mathcal{N}(y_n; x_n^T w, \sigma^2), \tag{6}$$

where σ is a (fixed) standard deviation.

It is interesting to note that this likelihood is maximized by the least squares solution; this is a generalization of standard linear regression

Bayesian Linear Regression: Posterior

We can now combine these to give the posterior using Bayes' rule:

$$p(w|\mathbf{x},\mathbf{y}) \propto p(w)p(\mathbf{y}|\mathbf{x},w) \tag{7}$$

$$= \mathcal{N}(w;0,C)\prod_{n=1}^{N}\mathcal{N}(y_n;x_n^T w,\sigma^2) \tag{8}$$

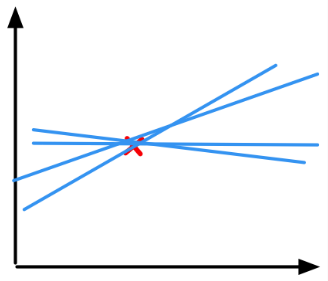
We omit the necessary algebra (see C M Bishop. *Pattern recognition and machine learning*. 2006, Chapter 3), but it is reasonably straightforward to show that

$$p(w|\mathbf{x},\mathbf{y}) = \mathcal{N}\left(w;m,S\right) \tag{9}$$

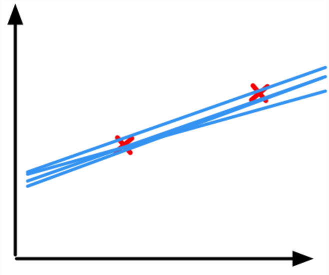
$$\text{where} \quad m = S^{-1}\mathbf{x}^T\mathbf{y}/\sigma^2 \quad \text{and} \quad S = \left(C^{-1} + \frac{\mathbf{x}^T\mathbf{x}}{\sigma^2}\right)^{-1}.$$

Bayesian Linear Regression: Posterior (2)

Note here that the fact the prior and posterior share the same form is highly special case. This is known as a conjugate distribution and it is why we were able to find an analytic solution for the posterior.

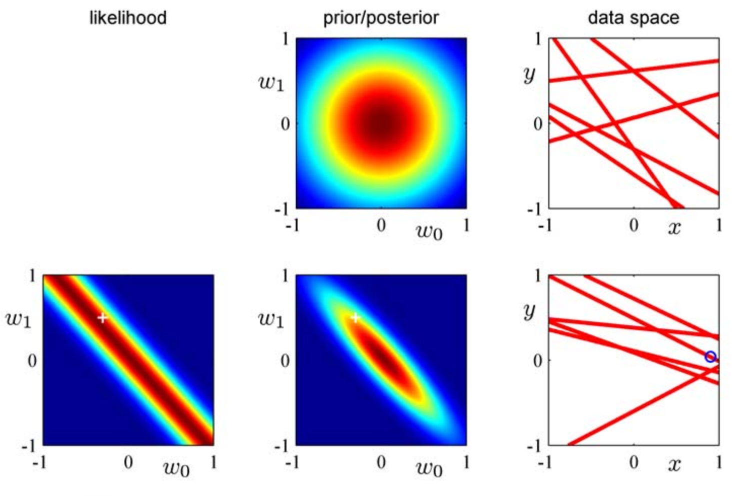


Posterior after 1 observation



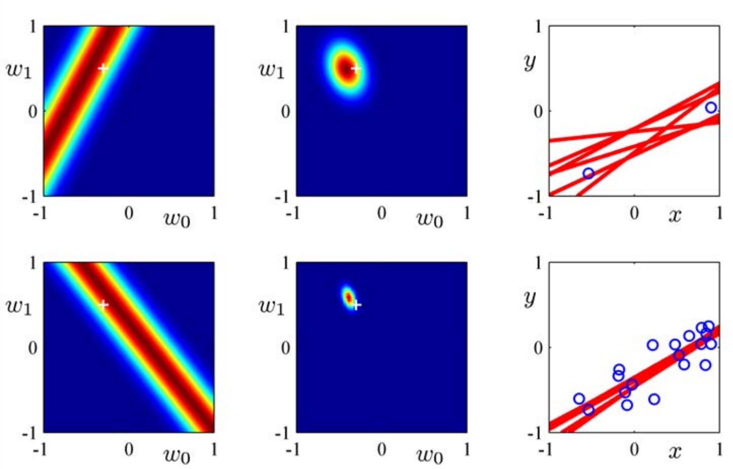
Posterior after 2 observations

Bayesian Linear Regression: Posterior (3)



[1] C M Bishop. *Pattern recognition and machine learning*. 2006.

Bayesian Linear Regression: Posterior (3)



[1] Bishop, *Pattern recognition and machine learning*.

Bayesian Linear Regression: Posterior Predictive

Given this posterior, we can now calculate the posterior predictive as follows

$$\begin{aligned} p(\tilde{y}|\tilde{x}, \mathbf{x}, \mathbf{y}) &= \int p(\tilde{y}|\tilde{x}, w) p(w|\mathbf{x}, \mathbf{y}) dw \qquad (10) \\ &= \int \mathcal{N}(\tilde{y}; \tilde{x}^T w, \sigma^2)\, \mathcal{N}\,(w; m, S)\, dw \\ &= \mathcal{N}\left(\tilde{y}\ ;\ \tilde{x}^T m,\ \left(\tilde{x}^T S^{-1} \tilde{x} + \frac{1}{\sigma^2}\right)^{-1}\right) \qquad (11) \end{aligned}$$

where the result is again a consequence of standard Gaussian identities and m and S are as before.

Bayesian Linear Regression: Posterior Predictive (2)

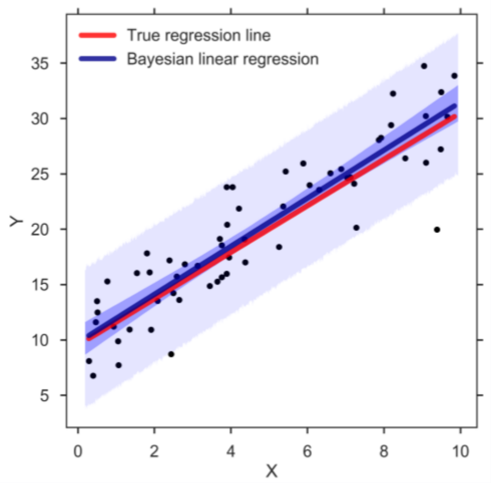
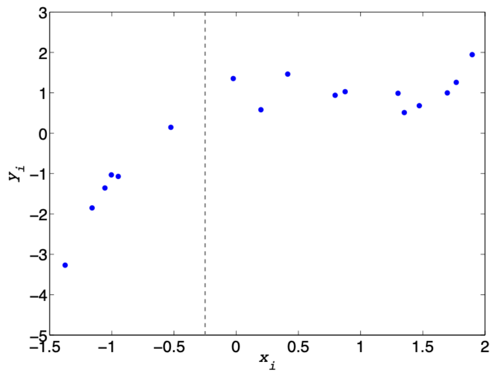


Image credit: https://www.dataminingapps.com/2017/09/simple-linear-regression-do-it-the-bayesian-way/

Model Comparison: What Makes a Good Model?

Example: Polynomial Regression

Imagine we are trying to fit this data:



Obviously linear regression is not appropriate here

This example is adapted from one by Carl Rasmussen: `http://mlg.eng.cam.ac.uk/teaching/4f13/1920/`

Example: Polynomial Regression (2)

A slightly more complicated approach would be to perform polynomial regression.

This is analogous to linear regression except that we now have (sticking to a one dimensional x for simplicity)

$$p(y_n|x_n, \mathbf{w}) = \mathcal{N}(y_n; w_0 + w_1 x_n + w_2 x_n^2 + \cdots + w_M x_n^M, \sigma^2), \quad (12)$$

where M is the degree of the polynomial, $\mathbf{w} = [w_1, w_2, \ldots, w_M]$, and we must now define a prior for each element of $\mathbf{w}$.

Example Polynomials for Each Degree

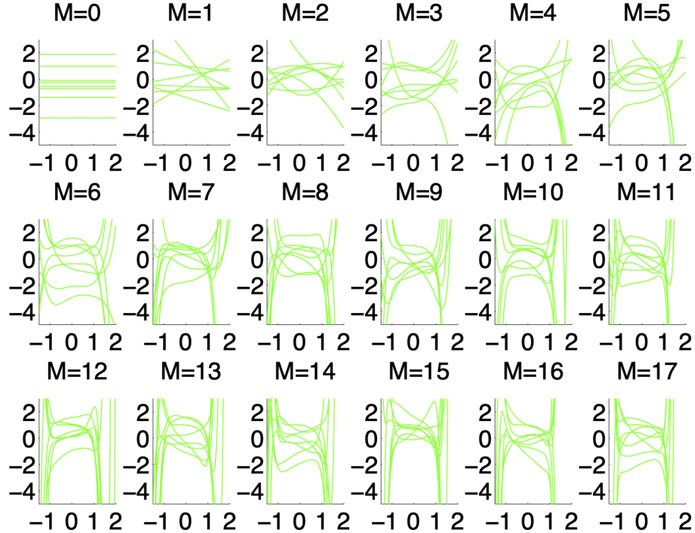


Image credit: Carl Rasmussen

Least Squares Estimate for Each Degree

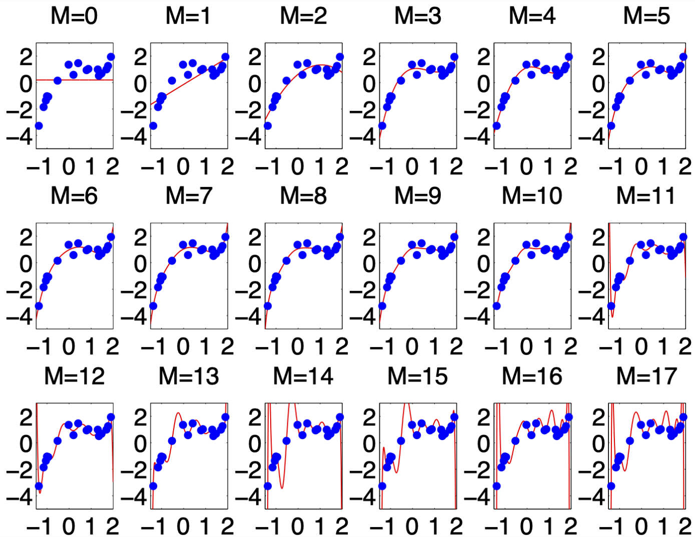


Image credit: Carl Rasmussen

Perfectly Matching the Data is Not Enough

With a high degree polynomial we can perfectly fit the data (i.e. achieve zero loss by passing through each point)...

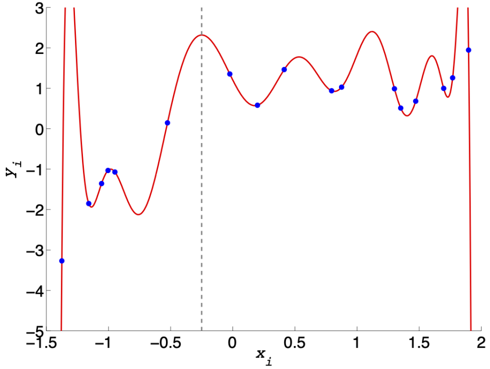


Image credit: Carl Rasmussen

Perfectly Matching the Data is Not Enough

With a high degree polynomial we can perfectly fit the data (i.e. achieve zero loss by passing through each point)...

...but the predictive power of this model is poor: we have **overfit**

We can actually have an arbitrarily large error at an unseen point (e.g. $x = 0.25$) while perfectly matching the data

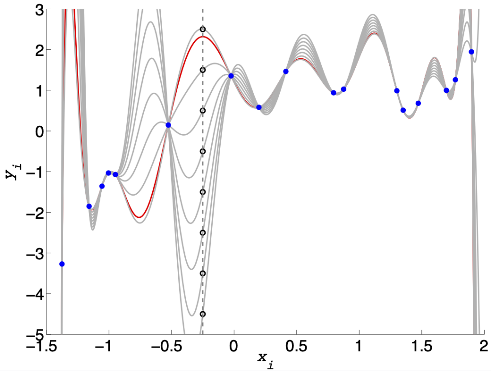


Image credit: Carl Rasmussen

What Went Wrong?

- We did not put a prior on the weights or regularize them in any other way
  - We implicitly took the following **maximum likelihood** solution that was very prone to overfitting:

$$\begin{aligned} p(y|x,\mathcal{D}) &= p(y|x,\theta^*) \\ \text{where} \quad \theta^* &= \arg\max_{\theta} p(\mathcal{D}|\theta) \end{aligned} \tag{13}$$

- We also did not careful consider if a high degree polynomial was actually a good model: having a good model requires more than just being able to fit the data
  - It also needs to **generalize** effectively to unseen samples

Marginal Likelihoods

- Let's revisit Bayes' rule and now condition on the model m:

$$p(\theta|\mathcal{D}, m) = \frac{p(\mathcal{D}|\theta, m)p(\theta|m)}{p(\mathcal{D}|m)} \tag{14}$$

- The marginal likelihood of a model $p(\mathcal{D}|m)$ represents the probability of the data under the model, averaging over all possible parameter values.
- It is crucial to deciding between models in Bayesian settings
  - A high marginal likelihood indicates a good model
  - For this reason, it also known as the **model evidence**
- The ratio of two marginal likelihoods, e.g. $p(\mathcal{D}|m_1)/p(\mathcal{D}|m_2)$, is known as a **Bayes factor** and is used to compare models

Marginal Likelihoods (2)

Why should we use the model evidence to compare models?

Apply Bayes rule to the models themselves:

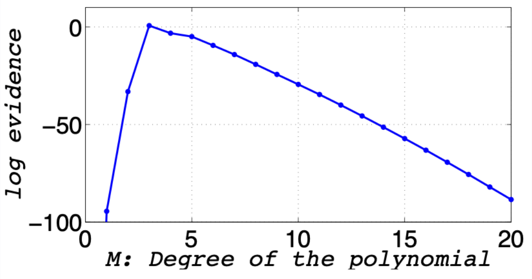
$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{p(\mathcal{D})} \tag{15}$$

This give us a direct relationship between the model evidence and the posterior probability of that model

If we a priori have no preference between models such that $p(m)$ is uniform, we even get that $p(m|\mathcal{D}) \propto p(\mathcal{D}|m)$

Marginal Likelihoods for Polynomial Regression

Returning to our polynomial regression problem and now taking a Bayesian approach with a Gaussian prior on the weights, we see that the model evidence prefers a degree of $M = 3$



This is a“sweet-spot”: complex enough to accurately match the data, simple enough to retain strong predictive power

Image credit: Carl Rasmussen

Bayesian Occam's Razor

- Occam's Razor states that if two explanations are able to explain a set of observations, the simpler one should be preferred.
- We can apply this in a Bayesian context by noting that the marginal likelihood is the probability that randomly selected parameters from the prior would generate $\mathcal{D}$
- Models that are too simple are unlikely to generate the observed dataset.
- Models that are too complex can generate many possible datasets, so again, they are unlikely to generate that particular dataset at random.

Bayesian Occam's Razor (2)

Imagine a hypothetical order on datasets where they get more complicated as we move away from the origin.

The model with highest evidence is the one that is powerful enough to explain that data but not anything more complicated.

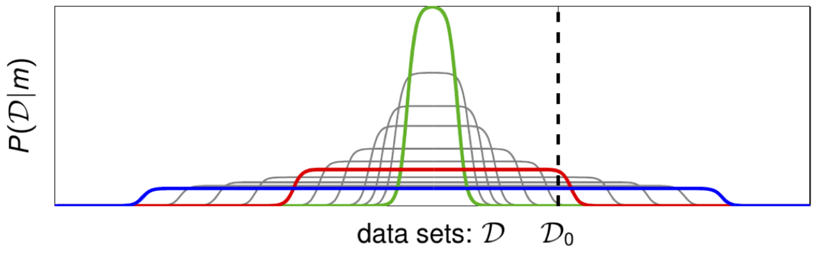


Image credit: Maneesh Sahani

Bayesian Occam's Razor (3)

We can visually see this effect by returning to our polynomial regression example (with different data this time):

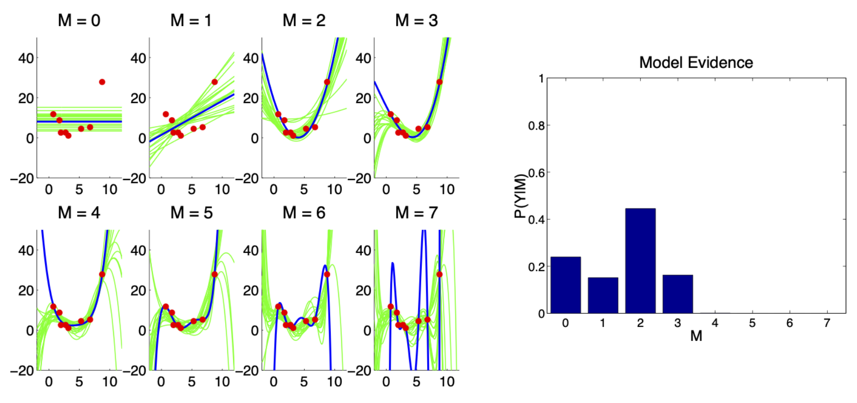


Image credit: Maneesh Sahani

Bayesian Model Averaging

Bayesian Model Averaging

Sometimes it is actually viable to avoid choosing between a set of models entirely by instead performing **Bayesian model averaging**

This involves being Bayesian about models themselves, namely marginalizing over them in the posterior predictive:

$$
\begin{aligned}
p(\mathcal{D}^*|\mathcal{D}) &= \iint p(\mathcal{D}^*|\theta, m)p(\theta|\mathcal{D}, m)p(m|\mathcal{D})d\theta dm && (16)\\
&= \iint p(\mathcal{D}^*|\theta, m)\frac{p(\mathcal{D}|\theta, m)p(\theta|m)}{p(\mathcal{D}|m)}\frac{p(\mathcal{D}|m)p(m)}{p(\mathcal{D})}d\theta dm && (17)\\
&= \iint p(\mathcal{D}^*|\theta, m)\frac{p(\mathcal{D}|\theta, m)p(\theta|m)p(m)}{p(\mathcal{D})}d\theta dm && (18)\\
&= \iint p(\mathcal{D}^*|\theta, m)p(\theta, m|\mathcal{D})d\theta dm && (19)
\end{aligned}
$$

where $p(m)$ is a prior on models and we see this is effectively equivalent to the standard posterior predictive for both $\{\theta, m\}$

Bayesian Model Averaging is not Model Combination

- It is important to note that Bayesian model averaging does not enrich the class of models themselves.
- Analogously to standard Bayesian inference, we are implicitly assuming **one** of the models has lead to the data: the averaging is over our own uncertainty, not a way of creating a more complex compound model
- Example: Bayesian decision trees are inherently less powerful than random forests. Given enough data the posterior collapses to a single tree that is itself usually a poor model

Further Reading

- Information on non-parametric models and Gaussian processes will be provided in the notes
- Bishop, *Pattern recognition and machine learning*, Chapters 1-3
- K P Murphy. *Machine learning: a probabilistic perspective*. 2012, Chapter 5
- D Barber. *Bayesian reasoning and machine learning*. 2012, Chapter 12
- T P Minka. “Bayesian model averaging is not model combination”. In: (2000)
- Zoubin Ghahramani on Bayesian machine learning (there are various alternative variations of this talk): `https://www.youtube.com/watch?v=y0FgHOQhG4w`
- Iain Murray on Probabilistic Modeling `https://www.youtube.com/watch?v=p0tvyVYAuW4`