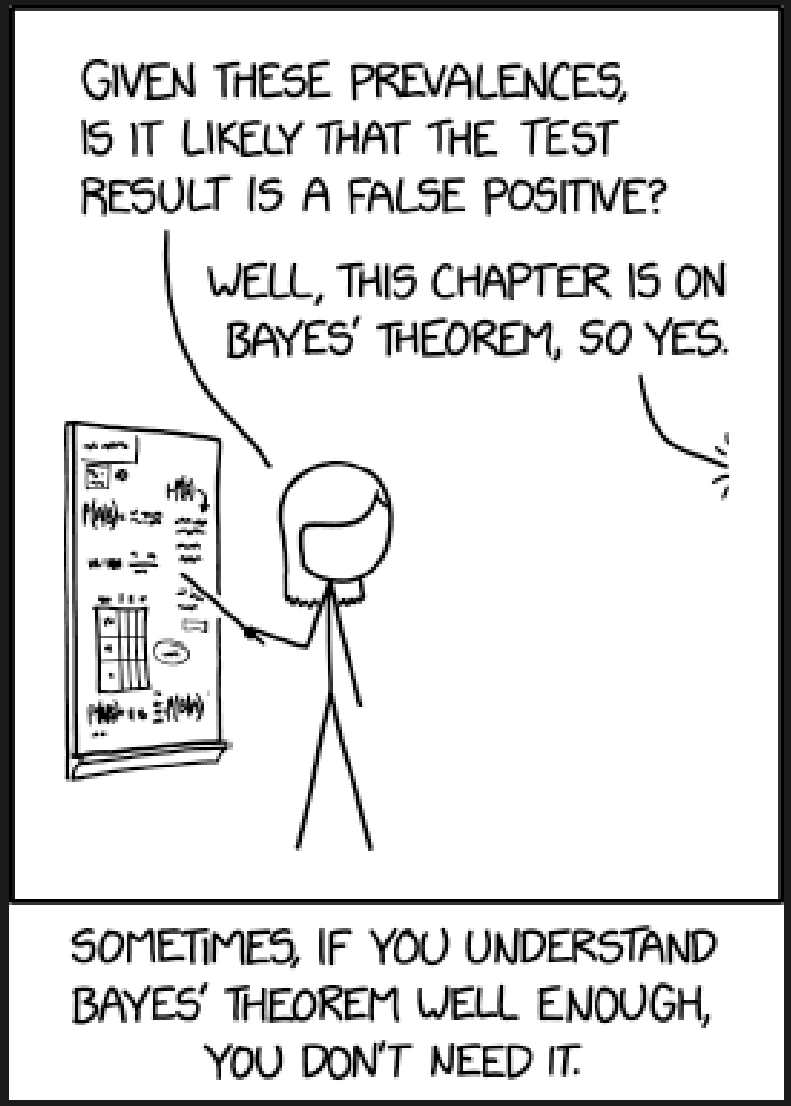
A lightning tour of Bayesian Statistics and Regularisation

Daniel Lawson — University of Bristol

Lecture 08.1 (v2.0.1)

In anticipation



Signposting

- Bayesian methodology is a huge and important area
- This is to give the background for:
  - Bayes Rule
  - Signposting Bayesian tools
  - Understanding Latent Dirichlet Allocation
  - Regularisation

Questions

- Can we still do inference when the posterior is intractable?
- Can we justify regularization?
- Are Bayesian methods slower than frequentist methods?
- Do we need to believe our prior?

A brief aside into Bayesian Modelling

- Bayesian Models are **generative**, that is, you can simulate data from them.
- They consist of:
  - a **prior** $\Pr(\theta)$, that is conceptualised as either a model, or as beliefs,
  - and the **likelihood** $\Pr(x|\theta)$, that depends on the data.
- The task is to integrate over the prior, to find the **posterior probability** using **Bayes' theorem**:

$$\Pr(\theta|x) = \frac{\Pr(x|\theta)\Pr(\theta)}{\Pr(x)}$$

- In general $\Pr(x)$ is hard to evaluate but there are methods to avoid doing this.

Example of Bayes Theorem

- One important application of Bayes' theorem is **False discovery**.
  - Imagine that we made a Bad-Guy-Detector (TM) which has a 99% chance of seeing a malicious attack if present ($\theta = 1$)...
  - But a 0.01% chance of declaring an attack when it isn't ($\theta = 0$).
  - Let p be the true frequency of malicious attacks.
  - If our BGD activates ($x = 1$), what is the probability of a true attack?
- Probability of the data: $\Pr(x = 1) = 0.99p + 0.0001(1 - p)$
- Probability of an attack: $\Pr(\theta = 1|x = 1) = 0.99p/\Pr(x = 1)$
- If $p = 0.001$ then $\Pr(\theta = 1|x = 1) \approx 0.9$
- If $p = 0.0001$ then $\Pr(\theta = 1|x = 1) \approx 0.5$
- If $p = 0.00001$ then $\Pr(\theta = 1|x = 1) \approx 0.09$
- If $p = 0.000001$ then $\Pr(\theta = 1|x = 1) \approx 0.001$

Etymology of Bayes: Conjugacy and tractability

- Bayesian Inference techniques can be used to **integrate out** model parameters:
- A **conjugate** model allows parameters to be integrated out analytically: i.e. you can compute $\Pr(x)$ and therefore $\Pr(\theta|x)$
- Monte-Carlo methods allow **sampling of posterior parameters** $\Pr(\theta|x)$ conditional on the data without ever evaluating $\Pr(x)$
- Some models are **doubly intractable**[1] meaning that you cannot compute $\Pr(x|\theta)$ and they cannot be sampled.
  - For example, Markov Random Fields.
  - Special methods are needed for them, for example, **Approximate Bayesian Computation**

[1] Murray, Ghahramani, and MacKay. "MCMC for doubly-intractable distributions." arXiv preprint arXiv:1206.6848 (2012).

Conjugate models

- Conjugate models take the form of a **known distribution** for the Prior, that can be updated through observations to the same distribution but with **new parameters**.
- Updating conjugate models with new data is straightforward: we can do it **online** by visiting each datapoint only once.
- We can also form a low-dimensional summary that captures everything about an observation.
- This means we can interpret the prior in terms of **pseudo observations**:
  - either data we have seen already,
  - or data we pretend to have seen in order to specify a prior distribution.
- The set of possible conjugate models is limited, though they can often be used as a part of a larger model.
  - For example, we might have a set of conjugate models to summarise several different data sources on a stream, which we then combine into a full, more costly model containing only a few non-conjugate parameters.

Conjugate model example

- Example: The Beta-Bernoulli model for binary outcomes.
  - In the Bernoulli model $p(x|p)$ we flip a (biased) coin x which is heads ($x = 1$) with some unknown probability p.
  - If we parameterise the prior $p(p) = \mathrm{Beta}(\alpha, \beta)$, with $\hat{p} = \alpha/(\alpha + \beta)$,
  - then after n observations $p(p|\{x\}) = \mathrm{Beta}(\alpha', \beta') = \mathrm{Beta}(\alpha + \sum_{i=1}^{n} x_i, \beta + (n - \sum_{i=1}^{n} x_i))$,
  - i.e. α was our prior number of successes (heads) and β our prior number of failures (tails).
- All discrete distributions with conjugate priors have this interpretation!
- Continuous distributions also contain a concept of *the number of observations used to form the prior estimate*.
- There is a super useful list of conjugate priors and interpretations on the Conjugate Prior Wikipedia page!

Markov Chain Monte Carlo (MCMC)

- MCMC[2] allows sampling from a posterior when we can evaluate the likelihood and the prior at any parameter value, but not integrate it.
- It performs a search of parameter space, comparing the posterior at the current point to the posterior at a proposed point, taking into account the probability of moving between the points in either direction.
- (Somewhat surprisingly) the set of samples taken over many iterations resembles a random sample from the posterior.
- This can be used to make predictions, estimate parameters, etc, by averaging over the samples.
- It is relatively costly - the number of likelihood evaluations required to obtain convergence is hard to predict.
- It is often a relatively good search algorithm for hard posteriors! Though careful choice of proposals is then needed.

[2] e.g. Gamerman and Hedibert. Markov chain Monte Carlo: stochastic simulation for Bayesian inference.

Tools for Bayesian Modelling using MCMC

- MCMC is very popular because it is straightforward to implement many models using it.
- Some important tools for Bayesian Inference allow models to be specified, an automatically do the inference for you using MCMC:
  - OpenBUGS (http://openbugs.net/w/FrontPage)
  - JAGS (http://mcmc-jags.sourceforge.net/)
  - STAN (http://mc-stan.org/)
- STAN is the current darling because it uses a clever method to sample, called the "no U-turn sampler" (NUTS) which searches parameter space with **Hamiltonian Monte Carlo**, a method that gives the search "momentum".

Sequential Monte Carlo (SMC) for filtering problems

- Filters are a class of model that take a sample of parameters and move them (through some observed space such as time) to track a changing distribution, for example, estimates of where an object is over time.
- Hidden Markov Models (HMMs) do this analytically for discrete parameter spaces, where the observation is a random variable depending on the true state of a system.
- The Kalman Filter is famous as it can be solved analytically by tracking a Normal distribution estimate of the location.
- Sequential Monte Carlo is a tool for implementing a wide range of Bayesian models.
- It was pioneered[3] and been integrated into MCMC[4] in Bristol.

---

[3] Doucet, Godsill, and Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering." Statistics and computing 10.3 (2000): 197-208.

[4] Andrieu, Doucet, and Holenstein Particle Markov chain Monte Carlo methods

Approximate Bayesian Computation (ABC)

- ABC[5] allows inference when the Likelihood cannot be evaluated, either because it is too costly, or the model is not described in terms of probabilities.
- It works by:
  - Simulating data from a model,
  - Creating a set of summary statistics from the data,
  - Comparing the summary statistics of the simulated data to the real data,
  - Accepting parameters that generate sufficiently close data.
- ABC posteriors are sampled using rejection, MCMC, or SMC.
- Simulation Based Inference (SBI) exploits neural networks to map data to parameters.
- Both can be computationally costly unless the simulation is fast.

[5] Beaumont, Zhang, and Balding. "Approximate Bayesian computation in population genetics." Genetics 162.4 (2002): 2025-2035.

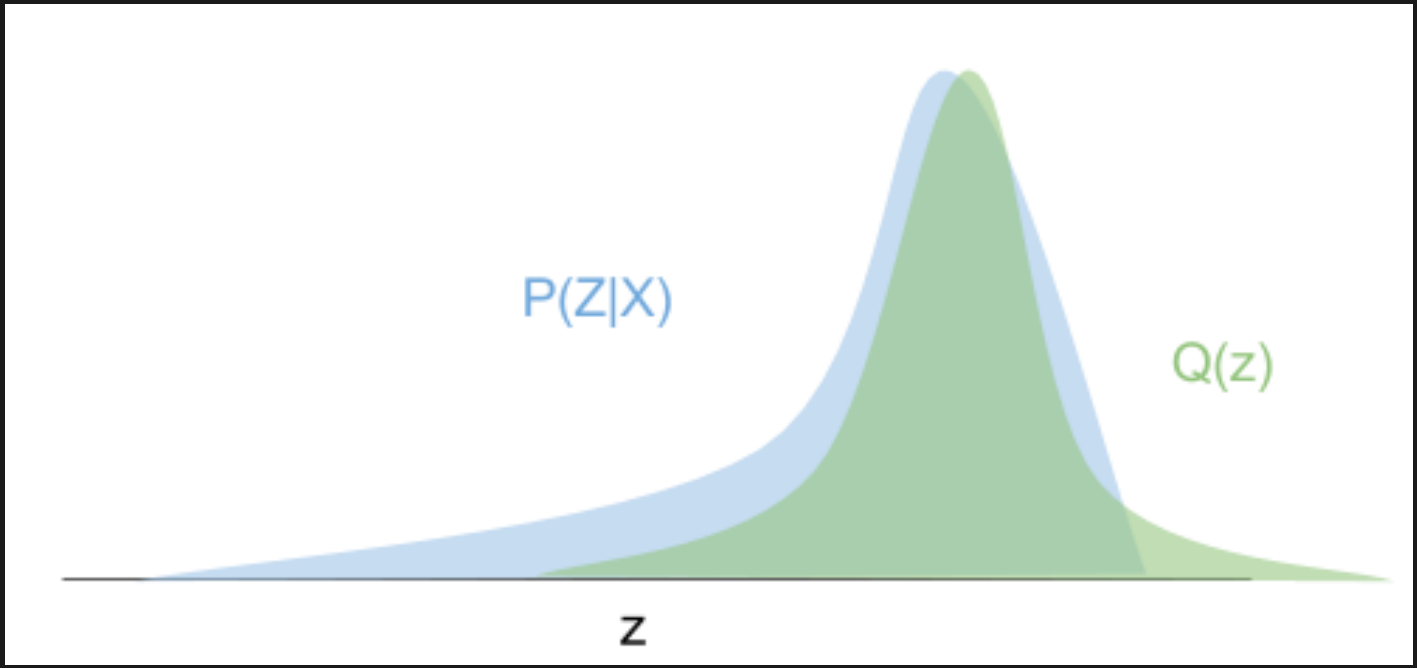
Bayesian Modelling in Machine Learning

- Machine Learning techniques need to be fast, so concentrate on **conjugate models**, or approximations that are the **nearest conjugate model**.
  - **Variational methods**[6] are extremely important for this.
  - The integration is turned into an optimisation problem, searching for the parameters that best approximate the whole posterior distribution.

[6] Blei and Jordan. "Variational inference for Dirichlet process mixtures." Bayesian analysis 1.1 (2006): 121-143.

Variational methods insight

- Seeking the distribution Q that best approximates the true distribution P, measured in "KL-Divergence"[7].



[7] http://blog.evjang.com/2016/08/variational-bayes.html

Motivating Regularisation and Smoothing

- Taking the **maximum likelihood** estimate can sometimes lead to problems, for example, if from n trials we observe zero successes, we estimate $\hat{p} = 0$ and hence place zero probability on observing a head in the future!
- Instead, it is good practice to assume that the whole sample space is plausible for future values, i.e. assume that our prior contains observations from every outcome.
  - Common to take 1 pseudo observation from every category, or 1 pseudo observation from the null, etc
  - Also reasonable to take "a small number" (0.01 often used) to provide non-zero mass to "unobserved events"
- In practice, this allows **regularised frequentist inference** by taking the maximum aposteriori (MAP) estimate of a Bayesian model
- Conjugacy is only required if we want an analytical solution. MAP estimates are very useful elsewhere, provided stable estimators exist.

Why regularise?

- The above interpretation makes it clear that Regularisation will change our estimate:
  - The first time a "new" type of observation is made, such as a new category or cluster;
  - When the number of pseudo observations is not small compared to the amount of data.
- It is therefore essential when:
  - Making **predictive distributions** allowing for the possibility that we have not yet learned everything,
  - The total number of training observations is "small".
- Regularisation is essential when $p > n$ where we have more parameters than data and therefore no power to estimate them all.

Regularisation models for regression

- In regression we minimise $(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$ with respect to β.
- Regression is typically regularised with either:
  - **Ridge** penalisation, by adding $\lambda_r(\beta - \mathbf{c})^T(\beta - \mathbf{c})$ to penalise towards c using second moments,
  - **Lasso** penalisation, by adding $\lambda_1\|\beta - \mathbf{c}\|$ to penalise towards c using first moments,
  - **ElasticNet** penalisation, which combines the above.
- These have **direct interpretations** in terms of a Bayesian model.
  - Ridge regression is assuming prior observations at $\mathbf{c}$ (with count a function of λ_r)
  - Lasso regression assumes that the prior is a Laplace distribution instead

Comments on regularisation

- Simple regularisation models can be represented as pseudo-observations. This is conceptually and practically convenient.
- Others cannot. They may enjoy other advantages, for example:
  - Coming from a **justifiable Bayesian prior**. For example, a hierarchical model assumes that there is a grand mean from which local clusters are sampled. Clusters are penalised towards the mean above them in the hierarchy.
  - Providing **desirable consequences**. For example, Lasso regression can set some coefficients to exactly zero, which is a valuable complexity reduction.
- Regularisation is **not Bayesian modelling**, even though it typically has an interpretation as a prior:
  - In Bayesian inference, we **integrate** over the prior to get a posterior **distribution**.
  - In MAP estimation and regularisation, we take the a point estimate.
- Variational inference attempts to integrate over the prior, by finding the closest fitting integrable distribution.

Reflection

- Are Bayesian approaches inherently slow?
- When might MAP estimation and full Bayesian inference produce different predictions?
- How have we encountered regularisation previously?
  - How does it relate to **non-parametric** models?
  - How does it relate to Random Forests, decision trees and other flexible predictors?
- When would we regularise vs cross-validate?
- Keep looking for regularisation as we move through the course, especially in flexible machine learning systems such as neural networks.

References

- There is a super useful list of conjugate priors and interpretations on the Conjugate Prior Wikipedia page!
- Methodology:
  - Gamerman and Hedibert. Markov chain Monte Carlo: stochastic simulation for Bayesian inference.
  - Doucet, Godsill, and Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering." Statistics and computing 10.3 (2000): 197-208.
  - Andrieu, Doucet, and Holenstein Particle Markov chain Monte Carlo methods
- ABC:
  - Murray, Ghahramani, and MacKay. "MCMC for doubly-intractable distributions." arXiv:1206.6848 (2012).
  - Beaumont, Zhang, and Balding. "Approximate Bayesian computation in population genetics." Genetics 162.4 (2002): 2025-2035.
- Variational Inference:
  - Blei and Jordan. "Variational inference for Dirichlet process mixtures", Bayesian analysis 1.1 (2006): 121-143.
  - A Beginner's Guide to Variational Methods, by Eric Jang.