

Portfolio Question 7.1.P - Bayes in Machine Learning

Daniel Lawson University of Bristol

Portfolio 7.1.P (v4.0.0)

Portfolio Question 7.1.P - Bayes in Machine Learning

- ▶ How is the Bayesian approach used in machine learning?
 - ▶ Contrast simple approaches where it can be used to motivate methods, to an area where it is taken seriously.

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**¹ are extremely important for this.
 - ▶ The integration is turned into an optimisation problem, searching for the parameters that best approximate the whole posterior distribution.

¹Blei and Jordan. "Variational inference for Dirichlet process mixtures." Bayesian analysis 1.1 (2006): 121-143.

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 a posteriori (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

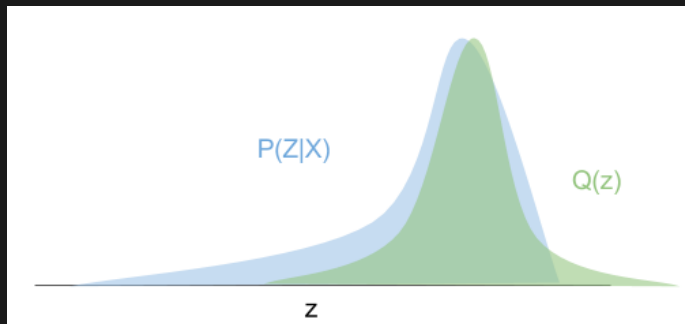
- ▶ In regression we minimise $(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$ with respect to β .
- ▶ Regression is typically regularised with either:
 - ▶ **Ridge** penalisation, by adding $\lambda_r(\beta - \mathbf{c})^T(\beta - \mathbf{c})$ to penalise towards $\mathbf{c}$ using second moments,
 - ▶ **Lasso** penalisation, by adding $\lambda_1\|\beta - \mathbf{c}\|$ to penalise towards $\mathbf{c}$ 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 useful 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 methods insight

- ▶ Seeking the distribution Q that best approximates the true distribution P , measured in “KL-Divergence”².
- ▶ Variational inference attempts to integrate over the prior, by finding the closest fitting integrable distribution.



²<http://blog.evjang.com/2016/08/variational-bayes.html>

Discussion

- ▶ Can we find examples of when the interpretation of regularisation doesn't make sense?
- ▶ Where is variational inference used?
 - ▶ save topic models for next session!
- ▶ What about “Bayesian Neural Networks”...?