CPSC 540: Machine Learning

Hierarchal Bayes

Mark Schmidt

University of British Columbia

Winter 2019

Last Time: Bayesian Predictions and Empirical Bayes

- We discussed making predictions using posterior predictive,

$$\hat{y} \in \underset{\tilde{y}}{\text{argmax}} \int_w p(\tilde{y} \mid \tilde{x}, w) p(w \mid X, y, \lambda) dw,$$

  which gives optimal predictions given your assumptions.

- We considered empirical Bayes (type II MLE),

$$\hat{\lambda} \in \underset{\lambda}{\text{argmax}}\, p(y \mid X, \lambda), \quad \text{where} \quad p(y \mid X, \lambda) = \int_w p(y \mid X, w) p(w \mid \lambda) dw,$$

  where we optimize marginal likelihood to select model and/or hyper-parameters.
  - Allows a huge number of hyper-parameters with less over-fitting than MLE.
  - Can use gradient descent to optimize continuous hyper-parameters.
  - Ratio of marginal likelihoods (Bayes factor) can be used for hypothesis testing.
  - In many settings, naturally encourages sparsity (in parameters, data, clusters, etc.).

Beta-Bernoulli Model

- Consider again a coin-flipping example with a Bernoulli variable,

$$x \sim \mathsf{Ber}(\theta).$$

- Last time we considered that either $\theta = 1$ or $\theta = 0.5$.
- Today: θ is a continuous variable coming from a beta distribution,

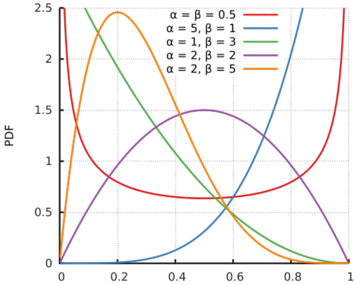
$$\theta \sim \mathcal{B}(\alpha, \beta).$$

- The parameters α and β of the prior are called hyper-parameters.
  - Similar to λ in regression, α and β are parameters of the prior.

Beta-Bernoulli Prior

Why the beta as a prior distribution?

- "It's a flexible distribution that includes uniform as special case".
- "It makes the integrals easy".



https://en.wikipedia.org/wiki/Beta_distribution

- Uniform distribution if $\alpha = 1$ and $\beta = 1$.
- "Laplace smoothing" corresponds to MAP with $\alpha = 2$ and $\beta = 2$.
  - Biased towards 0.5.

Beta-Bernoulli Posterior

- The PDF for the beta distribution has similar form to Bernoulli,

$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- Observing HTH under Bernoulli likelihood and beta prior gives posterior of

$$\begin{aligned} p(\theta \mid HTH, \alpha, \beta) &\propto p(HTH \mid \theta, \alpha, \beta) p(\theta \mid \alpha, \beta) \\ &\propto \left(\theta^2(1-\theta)^1\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) \\ &= \theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}. \end{aligned}$$

- Since proportionality ($\propto$) constant is unique for probabilities, posterior is a beta:

$$\theta \mid HTH, \alpha, \beta \sim \mathcal{B}(2+\alpha, 1+\beta).$$

- When the prior and posterior come from same family, it's called a conjugate prior.

Conjugate Priors

- Conjugate priors make Bayesian inference easier:

1. Posterior involves updating parameters of prior.
    - For Bernoulli-beta, if we observe h heads and t tails then posterior is $\mathcal{B}(\alpha + h, \beta + t)$.
    - Hyper-parameters α and β are "pseudo-counts" in our mind before we flip.
2. We can update posterior sequentially as data comes in.
    - For Bernoulli-beta, just update counts h and t.

Conjugate Priors

- Conjugate priors make Bayesian inference easier:

3. Marginal likelihood has closed-form as ratio of normalizing constants.
   - The beta distribution is written in terms of the beta function B,

$$p(\theta \mid \alpha, \beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{where} \quad B(\alpha,\beta) = \int_\theta \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta.$$

   and using the form of the posterior we have

$$p(HTH \mid \alpha, \beta) = \int_\theta \frac{1}{B(\alpha,\beta)}\theta^{(h+\alpha)-1}(1-\theta)^{(t+\beta)-1}d\theta = \frac{B(h+\alpha, t+\beta)}{B(\alpha,\beta)}.$$

   - Empirical Bayes (type II MLE) would optimize this in terms of α and β.

4. In many cases posterior predictive also has a nice form...

Bernoulli-Beta Posterior Predictive

If we observe 'HHH' then our different estimates are:

- MAP with uniform Beta(1,1) prior (maximum likelihood),

$$\hat{\theta} = \frac{(3+\alpha)-1}{(3+\alpha)+\beta-2} = \frac{3}{3} = 1.$$

- MAP Beta(2,2) prior (Laplace smoothing),

$$\hat{\theta} = \frac{(3+\alpha)-1}{(3+\alpha)+\beta-2} = \frac{4}{6} = \frac{2}{3}.$$

Bernoulli-Beta Posterior Predictive

If we observe 'HHH' then our different estimates are:

- Posterior predictive (Bayesian) with uniform Beta(1,1) prior,

$$\begin{aligned} p(H \mid HHH) &= \int_0^1 p(H \mid \theta)p(\theta \mid HHH)d\theta \\ &= \int_0^1 \mathrm{Ber}(H \mid \theta)\mathrm{Beta}(\theta \mid 3+\alpha, \beta)d\theta \\ &= \int_0^1 \theta\mathrm{Beta}(\theta \mid 3+\alpha, \beta)d\theta = \mathbb{E}[\theta] \\ &= \frac{4}{5}. \qquad \text{(mean of beta is } \alpha/(\alpha+\beta)\text{)} \end{aligned}$$

- Notice Laplace smoothing is not needed to avoid degeneracy under uniform prior.

Effect of Prior and Improper Priors

- We obtain different predictions under different priors:
  - $\mathcal{B}(3,3)$ prior is like seeing 3 heads and 3 tails (stronger prior towards 0.5),
    - For HHH, posterior predictive is 0.667.
  - $\mathcal{B}(100,1)$ prior is like seeing 100 heads and 1 tail (biased),
    - For HHH, posterior predictive is 0.990.
  - $\mathcal{B}(.01,.01)$ biases towards having unfair coin (head or tail),
    - For HHH, posterior predictive is 0.997.
    - Called "improper" prior (does not integrate to 1), but posterior can be "proper".
- We might hope to use an uninformative prior to not bias results.
  - But this is often hard/ambiguous/impossible to do (bonus slide).

Back to Conjugate Priors

- Basic idea of conjugate priors:

$$x \sim D(\theta), \quad \theta \sim P(\lambda) \quad \Rightarrow \quad \theta \mid x \sim P(\lambda').$$

- Beta-bernoulli example (beta is also conjugate for binomial and geometric):

$$x \sim \text{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta), \quad \Rightarrow \quad \theta \mid x \sim \mathcal{B}(\alpha', \beta'),$$

- Gaussian-Gaussian example:

$$x \sim \mathcal{N}(\mu, \Sigma), \quad \mu \sim \mathcal{N}(\mu_0, \Sigma_0), \quad \Rightarrow \quad \mu \mid x \sim \mathcal{N}(\mu', \Sigma'),$$

  and posterior predictive is also a Gaussian.
- If Σ is also a random variable:
  - Conjugate prior is normal-inverse-Wishart, posterior predictive is a student t.
- For the conjugate priors of many standard distributions, see:

  `https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions`

Back to Conjugate Priors

- Conjugate priors make things easy because we have closed-form posterior.
- Some “non-named” conjugate priors:
    - Discrete priors are “conjugate” to all likelihoods:
        - Posterior will be discrete, although it still might be NP-hard to use.
    - Mixtures of conjugate priors are also conjugate priors.
- Do conjugate priors always exist?
    - No, they only exist for exponential family likelihoods (next slides).
- Bayesian inference is ugly when you leave exponential family (e.g., student t).
    - Can use numerical integration for low-dimensional integrals.
    - For high-dimensional integrals, need Monte Carlo methods or variational inference.

Digression: Exponential Family

- Exponential family distributions can be written in the form

$$p(x \mid w) \propto h(x) \exp(w^T F(x)).$$

- We often have $h(x) = 1$, or an indicator that x satisfies constraints.
- $F(x)$ is called the sufficient statistics.
  - $F(x)$ tells us everything that is relevant about data x.
- If $F(x) = x$, we say that the w are cannonical parameters.
- Exponential family distributions can be derived from maximum entropy principle.
  - Distribution that is "most random" that agrees with the sufficient statistics $F(x)$.
  - Argument is based on "convex conjugate" of $-\log p$.

Digression: Bernoulli Distribution as Exponential Family

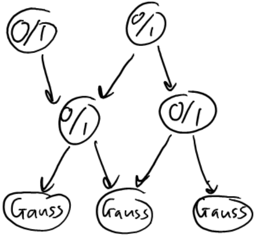
- We often define linear models by setting $w^T x^i$ equal to cannonical parameters.
- If we start with the Gaussian (fixed variance), we obtain least squares.
- For Bernoulli, the cannonical parameterization is in terms of "log-odds",

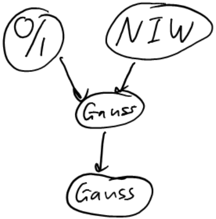
$$
\begin{aligned}
p(x \mid \theta) = \theta^x (1-\theta)^{1-x} &= \exp(\log(\theta^x (1-\theta)^{1-x})) \\
&= \exp(x \log \theta + (1-x)\log(1-\theta)) \\
&\propto \exp\left(x \log\left(\frac{\theta}{1-\theta}\right)\right).
\end{aligned}
$$

- Setting $w^T x^i = \log(y^i/(1-y^i))$ and solving for y^i yields logistic regression.
  - You can obtain regression models for other settings using this approach.

Conjugate Graphical Models

- DAG computations simplify if parents are conjugate to children.
- Examples:
  - Bernoulli child with Beta parent.
  - Gaussian belief networks.
  - Discrete DAG models.
  - Hybrid Gaussian/discrete, where discrete nodes can't have Gaussian parents.
  - Gaussian graphical model with normal-inverse-Wishart parents.





Outline

Hierarchical Bayesian Models

- Type II maximum likelihood is not really Bayesian:
  - We're dealing with w using the rules of probability.
  - But we're treating λ as a parameter, not a nuissance variable.
    - You could overfit λ.
- Hierarchical Bayesian models introduce a hyper-prior $p(\lambda \mid \gamma)$.
  - We can be "very Bayesian" and treat the hyper-parameter as a nuissance parameter.
- Now use Bayesian inference for dealing with λ:
  - Work with posterior over λ, $p(\lambda \mid X, y, \gamma)$, if integral over w is easy.
  - Or work with posterior over w and λ.
  - You could also consider a Bayes factor for comparing λ values:

$$p(\lambda_1 \mid X, y, \gamma)/p(\lambda_2 \mid X, y, \gamma),$$

which now account for belief in different hyper-parameter settings.

Model Selection and Averaging: Hyper-Parameters as Variables

- Bayesian model selection ("type II MAP"): maximizes hyper-parameter posterior,

$$\hat{\lambda} = \operatorname*{argmax}_{\lambda} p(\lambda \mid X, y, \gamma)$$
$$= \operatorname*{argmax}_{\lambda} p(y \mid X, \lambda) p(\lambda \mid \gamma),$$

  further taking us away from overfitting (thus allowing more complex models).
  - We could do the same thing to choose order of polynomial basis, σ in RBFs, etc.
- Bayesian model averaging considers posterior over hyper-parameters,

$$\hat{y}^i = \operatorname*{argmax}_{\hat{y}} \int_{\lambda} \int_{w} p(\hat{y} \mid \hat{x}^i, w) p(w, \lambda \mid X, y, \gamma) dw d\lambda.$$

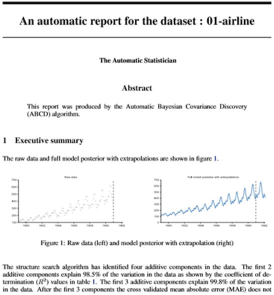
- Could maximize marginal likelihood of hyper-hyper-parameter γ, ("type III ML"),

$$\hat{\gamma} = \operatorname*{argmax}_{\gamma} p(y \mid X, \gamma) = \operatorname*{argmax}_{\gamma} \int_{\lambda} \int_{w} p(y \mid X, w) p(w \mid \lambda) p(\lambda \mid \gamma) dw d\lambda.$$

Application: Automated Statistician

- Hierarchical Bayes approach to regression:
  1. Put a hyper-prior over possible hyper-parameters.
  2. Use type II MAP to optimize hyper-parameters of your regression model.

- Can be viewed as an automatic statistician:
  `http://www.automaticstatistician.com/examples`



An automatic report for the dataset : 01-airline

The Automatic Statistician

Abstract

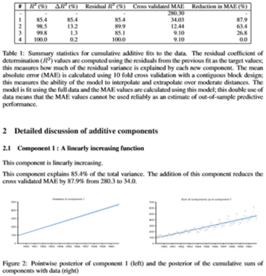
This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified four additive components in the data. The first 2 additive components explain 98.5% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 3 additive components explain 99.8% of the variation in the data. After the first 3 components the cross validated mean absolute error (MAE) does not



#	R^2 (%)	ΔR^2 (%)	Residual R^2 (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	280.30	-
1	85.4	85.4	85.4	34.03	87.9
2	98.5	13.2	89.9	12.44	63.4
3	99.8	1.3	85.1	9.10	26.8
4	100.0	0.2	100.0	9.10	0.0

Table 1: Summary statistics for cumulative additive fits to the data. The residual coefficient of determination (R^2) values are computed using the residuals from the previous fit as the target values; this measures how much of the residual variance is explained by each new component. The mean absolute error (MAE) is calculated using 10 fold cross validation with a contiguous block design; this measures the ability of the model to interpolate and extrapolate over moderate distances. The model is fit using the full data and the MAE values are calculated using this model; this double use of data means that the MAE values cannot be used reliably as an estimate of out-of-sample predictive performance.

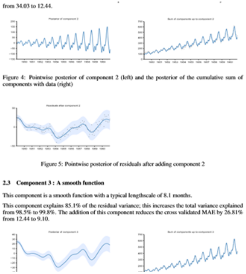
2 Detailed discussion of additive components

2.1 Component 1 : A linearly increasing function

This component is linearly increasing.

This component explains 85.4% of the total variance. The addition of this component reduces the cross validated MAE by 87.9% from 280.3 to 34.0.

Figure 2: Pointwise posterior of component 1 (left) and the posterior of the cumulative sum of components with data (right)



from 34.03 to 12.44.

Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Figure 5: Pointwise posterior of residuals after adding component 2

2.3 Component 3 : A smooth function

This component is a smooth function with a typical lengthscale of 8.1 months.

This component explains 85.1% of the residual variance; this increases the total variance explained from 98.5% to 99.8%. The addition of this component reduces the cross validated MAE by 26.81% from 12.44 to 9.10.

Discussion of Hierarchical Bayes

- "Super Bayesian" approach:
  - Go up the hierarchy until model includes all assumptions about the world.
  - Some people try to do this, and have argued that this may be how humans reason.
- Key advantage:
  - Mathematically simple to know what to do as you go up the hierarchy:
    - Same math for w, z, λ, γ, and so on (all are nuissance parameters).
- Key disadvantages:
  - It can be hard to exactly encode your prior beliefs.
  - The integrals get ugly very quickly.

Hierarchical Bayes as a Graphical Model

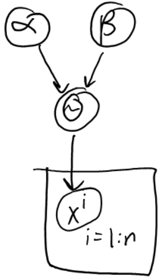
- Let x^i be a binary variable, representing if treatment works on patient i,

$$x^i \sim \text{Ber}(\theta).$$

- As before, let's assume that θ comes from a beta distribution,

$$\theta \sim \mathcal{B}(\alpha, \beta).$$

- We can visualize this as a graphical model:

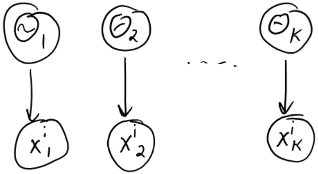


Hierarchical Bayes for Non-IID Data

- Now let x^i represent if treatment works on patient i in hospital j.
- Let's assume that treatment depends on hospital,

$$x_j^i \sim \text{Ber}(\theta_j).$$

- So the x_j^i are only IID given the hospital.



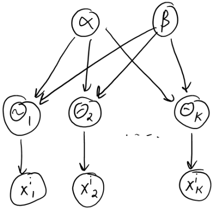
- Problem: we may not have a lot of data for each hospital.
  - Can we use data from one hospital to learn about others?
  - Can we say anything about a hospital with no data?

Hierarchical Bayes for Non-IID Data

- Common approach: assume the θ_j are drawn from common prior,

$$\theta_j \sim \mathcal{B}(\alpha, \beta).$$

- This introduces dependency between parameters at different hospitals:

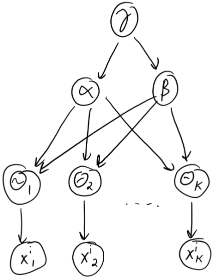


- But, if you fix α and β then you can't learn across hospitals:
    - The θ_j and d-separated given α and β.

- Type II MLE would optimize α and β given non-IID data.

Hierarchical Bayes for Non-IID Data

- Consider treating α and β as random variables and using a hyperprior:



- Now there is a dependency between the different θ_j (for unknown α and β).

- Now you can combine the non-IID data across different hospitals.
  - Data-rich hospitals inform posterior for data-poor hospitals.
  - You even consider the posterior for new hospitals with no data.

Summary

- Conjugate priors are priors that lead to posteriors of the same form.
  - They make Bayesian inference much easier.
- Exponential family distributions are the only distributions with conjugate priors.
- Hierarchical Bayes goes even more Bayesian with prior on hyper-parameters.
  - Leads to Bayesian model selection and Bayesian model averaging.
- Relaxing IID assumption with hierarchical Bayes.
- Next time: modeling cancer mutation signatures.

Uninformative Priors and Jeffreys Prior

- We might want to use an uninformative prior to not bias results.
  - But this is often hard/impossible to do.
- We might think the uniform distribution, $\mathcal{B}(1,1)$, is uninformative.
  - But posterior will be biased towards 0.5 compared to MLE.
  - And if you re-parameterize distribution it won't stay uniform.
- We might think to use "pseudo-count" of 0, $\mathcal{B}(0,0)$, as uninformative.
  - But posterior isn't a probability until we see at least one head and one tail.
- Some argue that the "correct" uninformative prior is $\mathcal{B}(0.5,0.5)$.
  - This prior is invariant to the parameterization, which is called a Jeffreys prior.