CPSC 540: Machine Learning

Hierarchal Bayes

Mark Schmidt

University of British Columbia

Winter 2018

Last Time: Bayesian Statistics

- For most of the course, we considered MAP estimation:

$$\hat{w} \in \underset{w}{\text{argmax}}\, p(w \mid X, y) \qquad \text{(train)}$$
$$\hat{y} \in \underset{\tilde{y}}{\text{argmax}}\, p(\tilde{y} \mid \tilde{x}, \hat{w}) \qquad \text{(test).}$$

- But w was random: I have no justification to only base decision on $\hat{w}$.
  - Ignores other reasonable values of w that could make opposite decision.
- Last time we introduced Bayesian approach:
  - Treat w as a random variable, and define probability over what we want given data:

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

  - Considers all the w, and weights their predictions by the posterior.
- Directly follows from rules of probability, and no separate training/testing.

Type II Maximum Likelihood for Regularization Parameter

- Maximum likelihood maximizes probability of data given parameters,

$$\hat{w} \in \underset{w}{\text{argmax}}\, p(y \mid X, w).$$

  - If we have a complicated model, this often overfits.
- Type II maximum likelihood maximizes probability of data given hyper-parameters,

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

  and the integral has closed-form solution if everything is Gaussian.
  - You can run gradient descent to choose λ.

- We are using the data to optimize the prior (empirical Bayes).
- Even if we have a complicated model, much less likely to overfit:
  - Complicated models need to integrate over many more alternative hypotheses.

Learning Principles

- Maximum likelihood:

$$\hat{w} \in \underset{w}{\text{argmax}}\, p(y \mid X, w) \qquad \hat{y} \in \underset{\tilde{y}}{\text{argmax}}\, p(\tilde{y} \mid \tilde{x}, \hat{w}).$$

- MAP:

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

  - Optimizing λ in this setting does not work: sets $\lambda = 0$.
- Bayesian (no "learning"):

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

- Type II maximum likelihood ("learn hyper-parameters"):

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

Type II Maximum Likelihood for Individual Regularization Parameter

- Consider having a hyper-parameter λ_j for each w_j,

$$y^i \sim \mathcal{N}(w^T x^i, \sigma^2 I), \quad w_j \sim \mathcal{N}(0, \lambda_j^{-1}).$$

- Too expensive for cross-validation, but type II MLE works.
  - You can do gradient descent to optimize the λ_j.

- Weird fact: this yields sparse solutions.
  - "Automatic relevance determination" (ARD)
  - Can send $\lambda_j \to \infty$, concentrating posterior for w_j at exactly 0.
    - It tries to "remove some of the integrals".
  - This is L2-regularization, but empirical Bayes naturally encourages sparsity.

- Non-convex and theory not well understood:
  - Tends to yield much sparser solutions than L1-regularization.

Type II Maximum Likelihood for Other Hyper-Parameters

- Consider also having a hyper-parameter σ_i for each i,

$$y^i \sim \mathcal{N}(w^T x^i, \sigma_i^2), \quad w_j \sim \mathcal{N}(0, \lambda_j^{-1}).$$

- You can also use type II MLE to optimize these values.
- The "automatic relevance determination" selects training examples ($\sigma_i \to \infty$).
  - This is like the support vectors in SVMs, but tends to be much more sparse.
- Type II MLE can also be used to learn kernel parameters like RBF variance.
  - Do gradient descent on the σ values in the Gaussian kernel.
- It will also do something sensible if you use it to choose number of clusters k.
  - Or number of states in hidden Markov model, number of latent factors in PCA, etc.
- Bonus slides: Bayesian feature selection gives probability that w_j is non-zero.
  - Posterior is much more informative than standard sparse MAP methods.

Outline

Beta-Bernoulli Model

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

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

- Last time we considered that either $\theta = 1$ or $\theta = 0.5$.

- Today: θ is a continuous variable coming from a beta distribution,

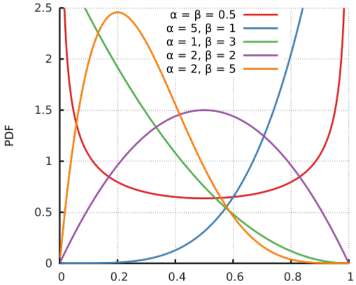
$$\theta \sim \mathcal{B}(\alpha, \beta).$$

- The parameters α and β of the prior are called hyper-parameters.
  - Similar to λ in regression, these 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 $\beta = 2$.

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

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

- So posterior is a beta distribution,

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

- 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:

- Maximum likelihood:
$$\hat{\theta} = \frac{n_H}{n} = \frac{3}{3} = 1.$$
- MAP with uniform Beta(1,1) prior,
$$\hat{\theta} = \frac{(3+\alpha)-1}{(3+\alpha)+\beta-2} = \frac{3}{3} = 1.$$
- Posterior predictive with uniform Beta(1,1) prior,
$$\begin{aligned} p(H \mid HHH) &= \int_0^1 p(H \mid \theta) p(\theta \mid HHH) d\theta \\ &= \int_0^1 \mathrm{Ber}(H \mid \theta)\mathrm{Beta}(\theta \mid 3+\alpha, \beta) d\theta \\ &= \int_0^1 \theta \mathrm{Beta}(\theta \mid 3+\alpha, \beta) d\theta = \mathbb{E}[\theta] \\ &= \frac{4}{5}. && \text{(using mean of beta formula)} \end{aligned}$$

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 uniform prior),
    - 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:

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

- 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.
- Two notable types of 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.
- 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 $h(x) = 1$, and $F(x)$ 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 the 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

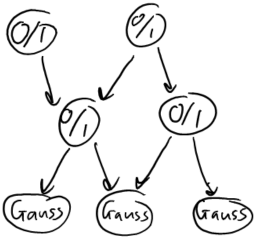
- We often define linear models by setting $w^T x^i$ equal to cannonical parameters.
- If we start with the Gaussian (fixed variance), we obtain least squares.
- For Bernoulli, the cannonical parameterization is in terms of "log-odds",

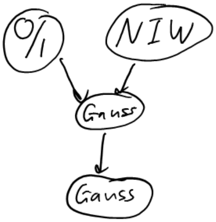
$$
\begin{aligned}
p(x \mid \theta) = \theta^x (1-\theta)^{1-x} &= \exp(\log(\theta^x (1-\theta)^{1-x})) \\
&= \exp(x \log \theta + (1-x) \log(1-\theta)) \\
&\propto \exp\left(x \log\left(\frac{\theta}{1-\theta}\right)\right).
\end{aligned}
$$

- 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 nuisance 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 nuisance parameter.
- Now use Bayesian inference for dealing with λ:
  - Work with posterior over λ, $p(\lambda \mid X, y, \gamma)$, or 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.

Bayesian Model Selection and Averaging

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

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

  which further takes 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,

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

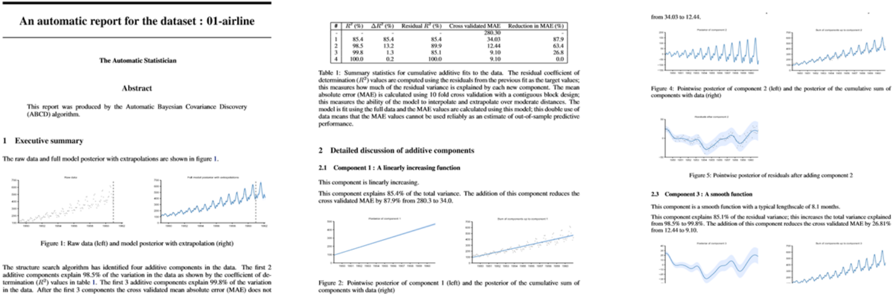
- Could maximize marginal likelihood of hyper-hyper-parameter γ, ("type III ML"),

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

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 nuisance parameters).
- Key disadvantages:
  - It can be hard to exactly encode your prior beliefs.
  - The integrals get ugly very quickly.

Summary

- Empirical Bayes optimizes marginal likelihood to set hyper-parameters:
  - Allows tuning a large number of hyper-parameters.
  - Bayesian Occam's razor: naturally encourages sparsity and simplicity.
- Conjugate priors are priors that lead to posteriors in the same family.
  - 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.
- 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.
- 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.

Gradient on Validation/Cross-Validation Error

- It's also possible to do gradient descent on λ to optimize validation/cross-validation error of model fit on the training data.

- For L2-regularized least squares, define $w(\lambda) = (X^TX + \lambda I)^{-1}X^Ty$.

- You can use chain rule to get derivative of validation error E_{valid} with respect to λ:

$$\frac{d}{d\lambda}E_{\text{valid}}(w(\lambda)) = E'_{\text{valid}}(w(\lambda))w'(\lambda).$$

- For more complicated models, you canuse total derivative to get gradient with respect to λ in terms of gradient/Hessian with respect to w.

- However, this is often more sensitive to over-fitting than empirical Bayes approach.

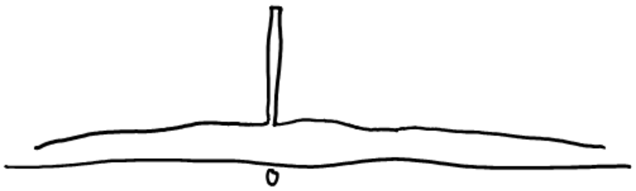
Bayesian Feature Selection

- Classic feature selection methods don't work when $d >> n$:
  - AIC, BIC, Mallow's, adjusted-R^2, and L1-regularization return very different results.
- Here maybe all we can hope for is posterior probability of $w_j = 0$.
  - Consider all models, and weight by posterior the ones where $w_j = 0$.
- If we fix λ and use L1-regularization, posterior is not sparse.
  - Probability that a variable is exactly 0 is zero.
  - L1-regularization only leads to sparse MAP, not sparse posterior.

Bayesian Feature Selection

- Type II MLE gives sparsity because posterior variance goes to zero.
  - But this doesn't give probabiliy of being 0.

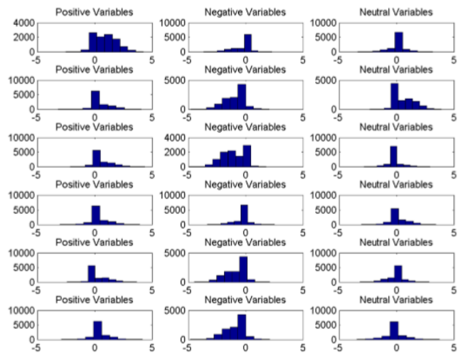
- We can encourage sparsity in Bayesian models using a spike and slab prior:



  - Mixture of Dirac delta function at 0 and another prior with non-zero variance.
  - Places non-zero posterior weight at exactly 0.
  - Posterior is still non-sparse, but answers the question:
    - "What is the probability that variable is non-zero"?

Bayesian Feature Selection

- Monte Carlo samples of w_j for 18 features when classifying '2' vs. '3':
  - Requires "trans-dimensional" MCMC since dimension of w is changing.



- "Positive" variables had $w_j > 0$ when fit with L1-regularization.
- "Negative" variables had $w_j < 0$ when fit with L1-regularization.
- "Neutral' variables had $w_j = 0$ when fit with L1-regularization.