CPSC 540: Machine Learning
Hierarchical Bayes

Mark Schmidt
University of British Columbia

Winter 2018
Last Time: Bayesian Statistics

- For most of the course, we considered **MAP estimation**:
  \[
  \hat{w} \in \arg\max_w p(w \mid X, y) \quad \text{(train)}
  \]
  \[
  \hat{y} \in \arg\max_{\tilde{y}} p(\tilde{y} \mid \tilde{x}, \hat{w}) \quad \text{(test)}.\]

- 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:
    \[
    \hat{y} \in \arg\max_{\tilde{y}} p(\tilde{y} \mid \tilde{x}, X, y)
    \]
    \[
    \equiv \arg\max_{\tilde{y}} \int_w p(\tilde{y} \mid \tilde{x}, w)p(w \mid X, y)dw.
    \]
  - 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,
  \[
  \hat{w} \in \arg\max_w p(y | X, w).
  \]

- If we have a complicated model, this often **overfits**.

- **Type II maximum likelihood** maximizes probability of data given hyper-parameters,
  \[
  \hat{\lambda} \in \arg\max_\lambda p(y | X, \lambda), \quad \text{where} \quad p(y | X, \lambda) = \int_w p(y | X, w)p(w | \lambda)dw,
  \]
  and the integral has closed-form solution if everything is Gaussian.
  - You can run gradient descent to choose \( \lambda \).

- 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:**
  \[ \hat{w} \in \arg\max_w p(y \mid X, w) \quad \hat{y} \in \arg\max_{\tilde{y}} p(\tilde{y} \mid \tilde{x}, \hat{w}). \]

- **MAP:**
  \[ \hat{w} \in \arg\max_w p(w \mid X, y, \lambda) \quad \hat{y} \in \arg\max_{\tilde{y}} p(\tilde{y} \mid \tilde{x}, \hat{w}). \]

  Optimizing \( \lambda \) in this setting does not work: sets \( \lambda = 0 \).

- **Bayesian (no “learning”):**
  \[ \hat{y} \in \arg\max_{\tilde{y}} \int_w p(\tilde{y} \mid \tilde{x}, w)p(w \mid X, y, \lambda)dw. \]

- **Type II maximum likelihood ( “learn hyper-parameters”):**
  \[ \hat{\lambda} \in \arg\max_{\lambda} p(y \mid X, \lambda) \quad \hat{y} \in \arg\max_{\tilde{y}} \int_w p(\tilde{y} \mid \tilde{x}, w)p(w \mid X, y, \hat{\lambda})dw. \]
Type II Maximum Likelihood for Individual Regularization Parameter

- Consider having a hyper-parameter $\lambda_j$ 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 $\lambda_j$.

- 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 \( \sigma_i \) 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 \( \sigma \) 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

1. Conjugate Priors

2. Hierarchical Bayes
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: \( \theta \) is a continuous variable coming from a beta distribution,
  \[ \theta \sim \mathcal{B}(\alpha, \beta). \]

- The parameters \( \alpha \) and \( \beta \) of the prior are called hyper-parameters.
  - Similar to \( \lambda \) 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”.

- Uniform distribution if $\alpha = 1$ and $\beta = 1$.
- “Laplace smoothing” corresponds to MAP with $\alpha = 2$ and $\beta = 2$.

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

https://en.wikipedia.org/wiki/Beta_distribution
Beta-Bernoulli Posterior

- The PDF for the beta distribution has similar form to Bernoulli,
  \[ p(\theta | \alpha, \beta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}. \]

- Observing HTH under Bernoulli likelihood and beta prior gives posterior of
  \[
  p(\theta | HTH, \alpha, \beta) \propto p(HTH | \theta, \alpha, \beta)p(\theta | \alpha, \beta) \\
  \propto \left( \theta^2(1 - \theta)^1\theta^{\alpha-1}(1 - \theta)^{\beta-1} \right) \\
  = \theta^{(2+\alpha)-1}(1 - \theta)^{(1+\beta)-1}.
  \]

- So posterior is a beta distribution,
  \[ \theta | HTH, \alpha, \beta \sim B(2 + \alpha, 1 + \beta). \]

- 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 $B(\alpha + h, \beta + t)$.
   - Hyper-parameters $\alpha$ and $\beta$ 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:
  - **Marginal likelihood** has closed-form as ratio of normalizing constants.
    - The beta distribution is written in terms of the beta function $B$,
      \[ p(\theta | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad \text{where} \quad B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} d\theta. \]
    - and using the form of the posterior we have
      \[ p(HTH | \alpha, \beta) = \int_0^1 \frac{1}{B(\alpha, \beta)} \theta^{(h+\alpha)-1} (1 - \theta)^{(t+\beta)-1} d\theta = \frac{B(h + \alpha, t + \beta)}{B(\alpha, \beta)}. \]
    - **Empirical Bayes** (type II MLE) would optimize this in terms of $\alpha$ and $\beta$.
  - 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,**
  \[
p(H \mid HHH) = \int_0^1 p(H \mid \theta)p(\theta \mid HHH)\,d\theta
  = \int_0^1 \text{Ber}(H \mid \theta)\text{Beta}(\theta \mid 3 + \alpha, \beta)\,d\theta
  = \int_0^1 \theta\text{Beta}(\theta \mid 3 + \alpha, \beta)\,d\theta = \mathbb{E}[\theta]
  = \frac{4}{5}. \]
  (using mean of beta formula)
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.

- $\mathcal{B}(0.01, 0.01)$ 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 \( \Sigma \) 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
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 \(- \log p\).
Conjugate Priors

Hierarchical Bayes

Digression: Bernoulli Distribution as Exponential Family

- We often define linear models by setting $w^T x^i$ equal to canonical parameters.

- If we start with the Gaussian (fixed variance), we obtain least squares.

- For Bernoulli, the canonical parameterization is in terms of “log-odds”:

  $$p(x | \theta) = \theta^x (1 - \theta)^{1-x} = \exp(\log(\theta^x (1 - \theta)^{1-x}))$$
  $$= \exp(x \log \theta + (1 - x) \log(1 - \theta))$$
  $$\propto \exp \left( x \log \left( \frac{\theta}{1 - \theta} \right) \right).$$

- 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

1. Conjugate Priors
2. Hierarchical Bayes
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 $\lambda$ as a parameter, not a nuisance variable.
    - You could overfit $\lambda$.

- Hierarchical Bayesian models introduce a hyper-prior $p(\lambda | \gamma)$.
  - We can be “very Bayesian” and treat the hyper-parameter as a nuisance parameter.

- Now use Bayesian inference for dealing with $\lambda$:
  - Work with posterior over $\lambda$, $p(\lambda | X, y, \gamma)$, or posterior over $w$ and $\lambda$.
  - You could also consider a Bayes factor for comparing $\lambda$ values:
    $$p(\lambda_1 | X, y, \gamma)/p(\lambda_2 | X, y, \gamma),$$

  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,
  \[
  \hat{\lambda} = \arg\max_{\lambda} p(\lambda \mid X, y, \gamma)
  \]
  \[
  = \arg\max_{\lambda} p(y \mid X, \lambda)p(\lambda \mid \gamma),
  \]
  which further takes us away from overfitting (thus allowing more complex models).
  - We could do the same thing to choose order of polynomial basis, $\sigma$ in RBFs, etc.

- **Bayesian model averaging** considers posterior over hyper-parameters,
  \[
  \hat{y}^i = \arg\max_{\hat{y}} \int_{\lambda} \int_{w} p(\hat{y} \mid \hat{x}^i, w)p(w, \lambda \mid X, y, \gamma)dwd\lambda.
  \]

- Could maximize **marginal likelihood of hyper-hyper-parameter $\gamma$**, ("type III ML"),
  \[
  \hat{\gamma} = \arg\max_{\gamma} p(y \mid X, \gamma) = \arg\max_{\gamma} \int_{\lambda} \int_{w} p(y \mid X, w)p(w \mid \lambda)p(\lambda \mid \gamma)dwd\lambda.
  \]
**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
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 $w$, $z$, $\lambda$, $\gamma$, 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.
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 $\lambda$ to optimize validation/cross-validation error of model fit on the training data.

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

- You can use chain rule to get derivative of validation error $E_{\text{valid}}$ with respect to $\lambda$:

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

- For more complicated models, you can use total derivative to get gradient with respect to $\lambda$ in terms of gradient/Hessian with respect to $w$.

- However, this is often more sensitive to over-fitting than empirical Bayes approach.
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 $\lambda$ 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 probability* 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”?
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.