CPSC 540: Machine Learning

Empirical Bayes, Hierarchical Bayes

Mark Schmidt

University of British Columbia

Winter 2017

Admin

- Assignment 5:
  - Due April 10.
- Project description on Piazza.
- Final details coming soon.
- Bonus lecture on April 10th (same time and place)?

Last Time: Bayesian Statistics

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

$$\hat{w} = \underset{w}{\operatorname{argmax}}\, p(w|X,y) \qquad \text{(train)}$$
$$\hat{y}^i = \underset{\hat{y}}{\operatorname{argmax}}\, p(\hat{y}|\hat{x}^i,\hat{w}) \qquad \text{(test)}.$$

- 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:

$$\begin{aligned}\hat{y}^i &= \underset{\hat{y}}{\operatorname{argmax}}\, p(\hat{y}|\hat{x}^i,X,y)\\ &= \underset{\hat{y}}{\operatorname{argmax}} \int_w p(\hat{y}|\hat{x}^i,w)p(w|X,y)dw.\end{aligned}$$

- Directly follows from rules of probability, and no separate training/testing.

7 Ingredients of Bayesian Inference

1. Likelihood $p(y|X, w)$.
2. Prior $p(w|\lambda)$.
3. Posterior $p(w|X, y, \lambda)$.
4. Predictive $p(\hat{y}|\hat{x}, w)$.
5. Posterior predictive $p(\hat{y}|\hat{x}, X, y, \lambda)$.
   - Probability of new data given old, integrating over parameters.
   - This tells us which prediction is most likely given data and prior.
6. Marginal likelihood $p(y|X, \lambda)$ (also called evidence).
   - Probability of seeing data given hyper-parameters.
   - We'll use this later for setting hyper-parameters.
7. Cost $C(\hat{y}|\tilde{y})$.
   - The penalty you pay for predicting $\hat{y}$ when it was really was $\tilde{y}$.
   - Leads to Bayesian decision theory: predict to minimize expected cost.

Decision Theory

- Consider a scenario where different predictions have different costs:

Predict / True	True "spam"	True "not spam"
Predict "spam"	0	100
Predict "not spam"	10	0

- Suppose we have found "good" parameters w.
- Instead of predicting most likely $\hat{y}$, we should minimize expected cost:

$$\begin{aligned}\mathbb{E}[\text{Cost}(\hat{y} = \text{“spam”})] = &\, p(\text{“spam”} \mid \hat{x}, w) C(\text{“spam”} \mid \text{“spam”}) \\ &+ p(\text{“not spam”} \mid \hat{x}, w) C(\text{“spam”} \mid \text{“not spam”}).\end{aligned}$$

- Consider a case where $p(\text{“spam”} \mid \hat{x}, w) > p(\text{“not spam”} \mid \hat{x}, w)$.
  - We might still predict "not spam" if expected cost is lower.

Bayesian Decision Theory

- Bayesian decision theory:
  - If we estimate w from data, we should use posterior predictive,

$$\begin{aligned}\mathbb{E}[\text{Cost}(\hat{y} = \text{"spam"})] &= p(\text{"spam"} \mid \hat{x}, X, y) C(\text{"spam"} \mid \text{"spam"}) \\ &+ p(\text{"not spam"} \mid \hat{x}, X, y) C(\text{"spam"} \mid \text{"not spam"}).\end{aligned}$$

  - Minimizing this expected cost is the optimal action.

- Note that there is a lot going on here:
  - Expected cost depends on cost and posterior predictive.
  - Posterior predictive depends on predictive and posterior
  - Posterior depends on likelihood and prior.

Outline

1. Empirical Bayes
2. Conjugate Priors
3. Hierarchical Bayes

Bayesian Linear Regression

- On Day 2, we argued that L2-regularized linear regression,

$$\operatorname*{argmin}_{w} \frac{1}{2\sigma^2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2,$$

corresponds to MAP estimation in the model

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

- By some tedious Gaussian identities, the posterior has the form

$$w|X, y \sim \mathcal{N}\left(\frac{1}{\sigma^2}A^{-1}X^T y, A^{-1}\right), \quad \text{with } A = \frac{1}{\sigma^2}X^T X + \lambda I.$$

- Notice that mean of posterior is the MAP estimate (not true in general).
- Bayesian perspective gives us variability in w and optimal predictions given prior.
- But it also gives different ways to choose λ and choose basis.

Learning the Prior from Data?

- Can we use the data to set the hyper-parameters?
- In theory: No!
  - It would not be a "prior".
  - It's no longer the right thing to do.

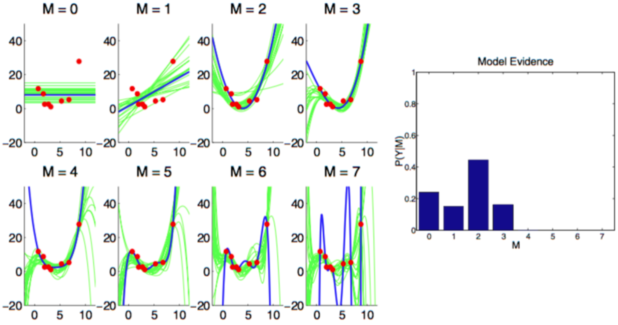
- In practice: Yes!
  - Approach 1: use a validation set or cross-validation as before.
  - Approach 2: optimize the marginal likelihood,

$$p(y|X,\lambda) = \int_w p(y|X,w)p(w|\lambda)dw.$$

  - Also called type II maximum likelihood or evidence maximization or empirical Bayes.

Type II Maximum Likelihood for Basis Parameter

- Consider polynomial basis, and treat degree M as a hyper-parameter:



http://www.cs.ubc.ca/~arnaud/stat535/slides5_revised.pdf

- Marginal likelihood (evidence) is highest for $M = 2$.
  - "Bayesian Occam's Razor": prefers simpler models that fit data well.
  - $p(y|X, \lambda)$ is small for $M = 7$, since 7-degree polynomials can fit many datasets.
  - Model selection criteria like BIC are approximations to marginal likelihood as $n \to \infty$.

Type II Maximum Likelihood for Regularization Parameter

- Maximum likelihood maximizes probability of data given parameters,

$$\hat{w} = \underset{w}{\mathrm{argmax}}\, p(y|X, w).$$

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

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

  and the integral has closed-form solution because posterior is Gaussian.
- We are using the data to optimize the prior.
- 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:

$$\hat{w} = \underset{w}{\text{argmax}}\, p(y|X, w) \qquad \hat{y}^i = \underset{\hat{y}}{\text{argmax}}\, p(\hat{y}|\hat{x}^i, \hat{w}).$$

- MAP:

$$\hat{w} = \underset{w}{\text{argmax}}\, p(w|X, y, \lambda) \qquad \hat{y}^i = \underset{\hat{y}}{\text{argmax}}\, p(\hat{y}|\hat{x}^i, \hat{w}).$$

- Optimizing λ in this setting does not work: sets $\lambda = 0$.
- Bayesian:

$$\hat{y}^i = \underset{\hat{y}}{\text{argmax}} \int_w p(\hat{y}|\hat{x}^i, w) p(w|X, y, \lambda) dw.$$

- Type II maximum likelihood:

$$\hat{\lambda} = \underset{\lambda}{\text{argmax}}\, p(y|, X, \lambda) \qquad \hat{y}^i = \underset{\hat{y}}{\text{argmax}} \int_w p(\hat{y}|\hat{x}^i, w) p(w|X, y, \hat{\lambda}) dw.$$

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 using log-marginal likelihood.

- Weird fact: yields sparse solutions (automatic relevance determination).
  - Can send $\lambda_j \to \infty$, concentrating posterior for w_j at 0.
  - 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.
  - This is like support vectors.
- Type II MLE can also be used to learn kernel parameters like RBF variance.
- Bonus slides: Bayesian feature selection gives probability that x_j is relevant.

Bayes Factors for Bayesian Hypothesis Testing

- Suppose we want to compare hypotheses:
  - E.g., L2-regularizer of λ_1 and L2-regularizer of λ_2.
- Bayes factor is ratio of marginal likelihoods,

$$\frac{p(y|X,\lambda_1)}{p(y|X,\lambda_2)}.$$

  - If very large then data is much more consistent with λ_1.
- A more direct method of hypothesis testing:
  - No need for null hypothesis, "power" of test, p-values, and so on.
  - But can only tell you which model is more likely, not whether any is correct.

- Last year from American Statistical Assocation:
  - “Statement on Statistical Significance and P-Values”:
    - http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108
- Bayes factors don’t solve problems with p-values.
  - But they give an alternative view, and make prior assumptions clear.
- Some notes on various issues associated with Bayes factors: http://www.aarondefazio.com/adefazio-bayesfactor-guide.pdf

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,

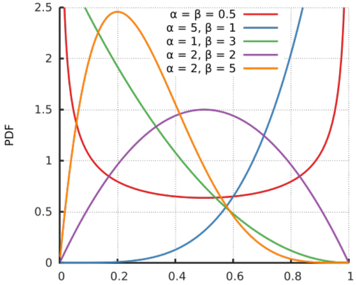
$$\theta \sim \mathcal{B}(\alpha, \beta).$$

- 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,

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

- Observing HTH under Bernoulli likelihood and beta prior then posterior is

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

- So posterior is a beta distribution,

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

- 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,

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

   and using the form of the posterior we have

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

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

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 other 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, only exist for exponential family likelihoods.
- Bayesian inference is ugly when you leave exponential family (e.g., student t).
  - Need Monte Carlo methods or variational inference.

Exponential Family

- Exponential family distributions can be written in the form

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

- 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 $-\log p$.

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",

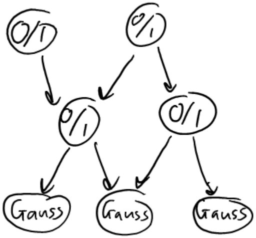
$$\begin{aligned} p(x|\theta) &= \theta^x (1-\theta)^{1-x} = \exp(\log(\theta^x (1-\theta)^{1-x})) \\ &= \exp(x \log \theta + (1-x)\log(1-\theta)) \\ &\propto \exp\left(x \log\left(\frac{\theta}{1-\theta}\right)\right). \end{aligned}$$

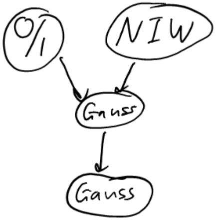
- Setting $w^T x^i = \log(y^i/(1-y^i))$ and solving for y^i yields logistic regression.

Conjugate Graphical Models

- DAG computations simplify if parents are conjugate to children.

- Examples:
  - Gaussian graphical models.
  - Discrete graphical 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 using a "point estimate" of λ.
- Hierarchical Bayesian models introduce a hyper-prior $p(\lambda|\gamma)$.
  - This is a "very Bayesian" model.
- Now use Bayesian inference for dealing with λ:
  - Work with posterior over λ, $p(\lambda|X, y, \gamma)$, or posterior over w and λ.
  - You could also consider a Bayes factor for comparing λ values:

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

Bayesian Model Selection and Averaging

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

$$\begin{aligned}\hat{\lambda} &= \underset{\lambda}{\operatorname{argmax}}\, p(\lambda|X,y,\gamma)\\ &= \underset{\lambda}{\operatorname{argmax}}\, p(y|X,\lambda)p(\lambda|\gamma),\end{aligned}$$

  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,

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

- We could also maximize marginal likelihood of γ, ("type III ML"),

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

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

Do we really need hyper-priors?

- In Bayesian statistics we work with posterior over parameters,

$$p(\theta|x,\alpha,\beta) = \frac{p(x|\theta)p(\theta|\alpha,\beta)}{p(x|\alpha,\beta)}.$$

- We discussed empirical Bayes, where you optimize prior using marginal likelihood,

$$\underset{\alpha,\beta}{\text{argmax}}\, p(x|\alpha,\beta) = \underset{\alpha,\beta}{\text{argmax}} \int_\theta p(x|\theta)p(\theta|\alpha,\beta)d\theta.$$

  - Can be used to optimize λ_j, polynomial degree, RBF σ_i, polynomial vs. RBF, etc.
- We also considered hierarchical Bayes, where you put a prior on the prior,

$$p(\alpha,\beta|x,\gamma) = \frac{p(x|\alpha,\beta)p(\alpha,\beta|\gamma)}{p(x|\gamma)}.$$

  - But is the hyper-prior really needed?

Hierarchical Bayes as Graphical Model

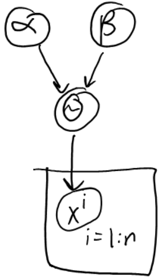
- Let x^i be a binary variable, representing if treatment works on patient i,

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

- 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:

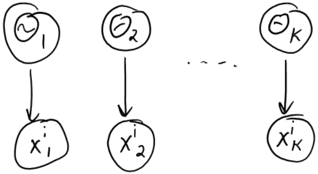


Hierarchical Bayes for Non-IID Data

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

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

- The x_j^i are IID given the hospital.



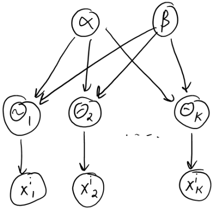
- But we may have more data for some hospitals than others:
  - Can we use data from one hospital to learn about others?
  - Can we say anything about a hospital with no data?

Hierarchical Bayes for Non-IID Data

- Common approach: assume θ_j drawn from common prior,

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

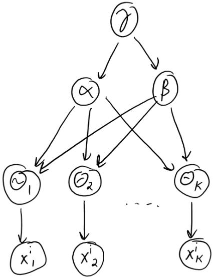
- This ties the parameters from the different hospitals together:



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

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.
- You 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.

Summary

- Marginal likelihood is probability seeing data given hyper-parameters.
- Empirical Bayes optimizes this 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.

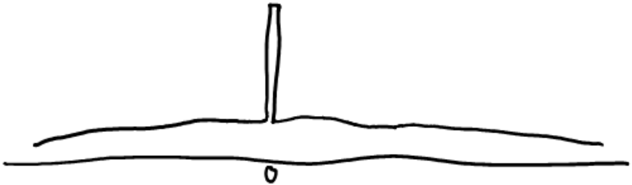
Bonus Slide: Bayesian Feature Selection

- Classic feature selection methods don't work whe $d >> n$:
  - 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 lead to sparse MAP, not sparse posterior.

Bonus Slide: 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"?

Bonus Slide: 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, $\mathcal{B}(0,0)$, are uninformative.
  - But posterior isn't a probability until we see at 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.