CPSC 540: Machine Learning

Empirical and Hierarchical Bayes

Mark Schmidt

University of British Columbia

Winter 2016

Admin

- Midterm:
  - Marks posted on UBC Connect.
- Assignment 5:
  - Out soon.
  - Due April 5th.
- Remaining topics:
  - More Bayesian stats, structured prediction, variational inference, deep learning.

Last Time: Bayesian Statistics

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

$$\hat{w} = \operatorname*{argmax}_{w} p(w|X,y) \qquad \text{(train)}$$

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

Last Time: Bayesian Statistics

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

$$\hat{w} = \underset{w}{\text{argmax}}\, p(w|X, y) \qquad \text{(train)}$$

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

- 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: Bayesian Statistics

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

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

- 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 week, we considered Bayesian approach:
    - Treat w as a random variable, and define probability over what we want given data:

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

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 week, we considered 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.

Beta-Bernoulli Model

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

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

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 let's view θ as a continuous random variable.

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 let's view θ as a continuous random variable.
- In particular, let's assume θ comes 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 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 let's view θ as a continuous random variable.
- In particular, let's assume θ comes 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.
- The PDF for the beta distribution has the form

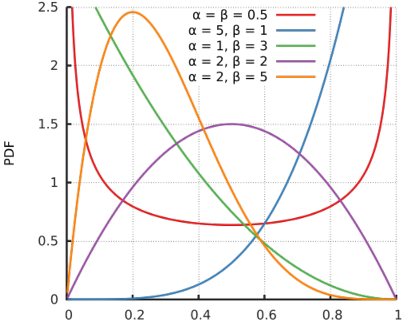
$$p(\theta|\alpha, \beta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where the beta function is $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$.

Beta-Bernoulli Model

Why the beta distribution?

- "It's a flexible distribution that includes uniform as special case".

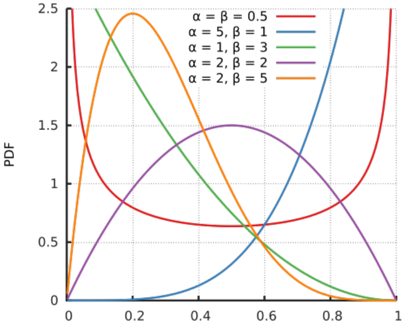


https://en.wikipedia.org/wiki/Beta_distribution

Beta-Bernoulli Model

Why the beta distribution?

- "It's a flexible distribution that includes uniform as special case".



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

- Uniform distribution if $\alpha = 1$ and $\beta = 1$.
- "It makes the integrals easy".

Ingredients of Bayesian Inference

1. Likelihood $p(x|\theta)$.
   - Probability of seeing data given parameters.

Ingredients of Bayesian Inference

1. Likelihood $p(x|\theta)$.
    - Probability of seeing data given parameters.
2. Prior $p(\theta|\alpha, \beta)$.
    - Belief that parameters are correct before we've seen data.

Ingredients of Bayesian Inference

1. Likelihood $p(x|\theta)$.
   - Probability of seeing data given parameters.
2. Prior $p(\theta|\alpha, \beta)$.
   - Belief that parameters are correct before we've seen data.
3. Posterior $p(\theta|x, \alpha, \beta)$.
   - Probability that parameters are correct after we've seen data.

Ingredients of Bayesian Inference

1. Likelihood $p(x|\theta)$.
   - Probability of seeing data given parameters.
2. Prior $p(\theta|\alpha, \beta)$.
   - Belief that parameters are correct before we've seen data.
3. Posterior $p(\theta|x, \alpha, \beta)$.
   - Probability that parameters are correct after we've seen data.
   - We won't use the MAP "point estimate", we want the whole distribution.

Ingredients of Bayesian Inference

1. Likelihood $p(x|\theta)$.
   - Probability of seeing data given parameters.
2. Prior $p(\theta|\alpha, \beta)$.
   - Belief that parameters are correct before we've seen data.
3. Posterior $p(\theta|x, \alpha, \beta)$.
   - Probability that parameters are correct after we've seen data.
   - We won't use the MAP "point estimate", we want the whole distribution.
4. Posterior predictive $p(\hat{x}|x, \alpha, \beta)$.
   - Probability of new data given old, integrating over parameters.
   - This tells us which prediction is most likely given data and prior.

Ingredients of Bayesian Inference

1. Likelihood $p(x|\theta)$.
   - Probability of seeing data given parameters.
2. Prior $p(\theta|\alpha, \beta)$.
   - Belief that parameters are correct before we've seen data.
3. Posterior $p(\theta|x, \alpha, \beta)$.
   - Probability that parameters are correct after we've seen data.
   - We won't use the MAP "point estimate", we want the whole distribution.
4. Posterior predictive $p(\hat{x}|x, \alpha, \beta)$.
   - Probability of new data given old, integrating over parameters.
   - This tells us which prediction is most likely given data and prior.
5. Marginal likelihood $p(x|\alpha, \beta)$ (also called evidence).
   - Probability of seeing data given hyper-parameters.

Ingredients of Bayesian Inference

1. Likelihood $p(x|\theta)$.
   - Probability of seeing data given parameters.
2. Prior $p(\theta|\alpha,\beta)$.
   - Belief that parameters are correct before we've seen data.
3. Posterior $p(\theta|x,\alpha,\beta)$.
   - Probability that parameters are correct after we've seen data.
   - We won't use the MAP "point estimate", we want the whole distribution.
4. Posterior predictive $p(\hat{x}|x,\alpha,\beta)$.
   - Probability of new data given old, integrating over parameters.
   - This tells us which prediction is most likely given data and prior.
5. Marginal likelihood $p(x|\alpha,\beta)$ (also called evidence).
   - Probability of seeing data given hyper-parameters.
6. We might also have a cost $g(\tilde{x}|\hat{x})$.
   - The penalty you pay for predicting $\hat{x}$ when it was really was $\tilde{x}$.

Ingredients of Bayesian Inference

1. Likelihood $p(x|\theta)$.
    - Probability of seeing data given parameters.
2. Prior $p(\theta|\alpha, \beta)$.
    - Belief that parameters are correct before we've seen data.
3. Posterior $p(\theta|x, \alpha, \beta)$.
    - Probability that parameters are correct after we've seen data.
    - We won't use the MAP "point estimate", we want the whole distribution.
4. Posterior predictive $p(\hat{x}|x, \alpha, \beta)$.
    - Probability of new data given old, integrating over parameters.
    - This tells us which prediction is most likely given data and prior.
5. Marginal likelihood $p(x|\alpha, \beta)$ (also called evidence).
    - Probability of seeing data given hyper-parameters.
6. We might also have a cost $g(\tilde{x}|\hat{x})$.
    - The penalty you pay for predicting $\hat{x}$ when it was really was $\tilde{x}$.
    - Leads to Bayesian decision theory.
    - Straightforward extension: predict to minimize expected cost.

Posterior and Marginal Likelihood

- Our model is: $x \sim \mathrm{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta)$.

Posterior and Marginal Likelihood

- Our model is: $x \sim \mathrm{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta)$.
- If we observe 'HTH' then our posterior distribution is

$$
\begin{aligned}
p(\theta|HTH, \alpha, \beta) &= \frac{p(HTH|\theta, \alpha, \beta)p(\theta|\alpha, \beta)}{p(HTH|\alpha, \beta)} && \text{(Bayes)} \\
&= \frac{\left(\theta^2(1-\theta)^1\right)\left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)}{p(HTH|\alpha, \beta)} && \text{(likelihood/prior)} \\
&= \frac{\frac{1}{B(\alpha,\beta)}\theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}}{p(HTH|\alpha, \beta)}.
\end{aligned}
$$

Posterior and Marginal Likelihood

- Our model is: $x \sim \text{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta)$.
- If we observe 'HTH' then our posterior distribution is

$$
\begin{aligned}
p(\theta | HTH, \alpha, \beta) &= \frac{p(HTH|\theta, \alpha, \beta)p(\theta|\alpha, \beta)}{p(HTH|\alpha, \beta)} && \text{(Bayes)} \\
&= \frac{\left(\theta^2(1-\theta)^1\right)\left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)}{p(HTH|\alpha, \beta)} && \text{(likelihood/prior)} \\
&= \frac{\frac{1}{B(\alpha,\beta)}\theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}}{p(HTH|\alpha, \beta)}.
\end{aligned}
$$

- Denominator is marginal likelihood,

$$
p(HTH|\alpha, \beta) = \int_\theta \frac{1}{B(\alpha, \beta)}\theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1} d\theta.
$$

Posterior and Marginal Likelihood

- Our model is: $x \sim \mathrm{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta)$.
- If we observe 'HTH' then our posterior distribution is

$$\begin{aligned} p(\theta|HTH, \alpha, \beta) &= \frac{p(HTH|\theta, \alpha, \beta)p(\theta|\alpha, \beta)}{p(HTH|\alpha, \beta)} && \text{(Bayes)} \\ &= \frac{\left(\theta^2(1-\theta)^1\right)\left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)}{p(HTH|\alpha, \beta)} && \text{(likelihood/prior)} \\ &= \frac{\frac{1}{B(\alpha,\beta)}\theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}}{p(HTH|\alpha, \beta)}. \end{aligned}$$

- Denominator is marginal likelihood,

$$p(HTH|\alpha, \beta) = \int_\theta \frac{1}{B(\alpha, \beta)}\theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}d\theta.$$

- Understanding Bayesian inference is much easier once you can notice that:
  - The posterior is a beta distribution and the marginal likelihood integral is trivial.

Posterior and Marginal Likelihood

- Given HTH, we've shown that posterior is

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

Posterior and Marginal Likelihood

- Given HTH, we've shown that posterior is

$$p(\theta|HTH,\alpha,\beta) = \frac{\frac{1}{B(\alpha,\beta)}\theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}}{p(HTH|\alpha,\beta)}$$
$$\propto \theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}.$$

- Consider a $\mathcal{B}(\alpha',\beta')$ distribution on θ with $\alpha' = 2+\alpha$ and $\beta' = 1+\beta$,

$$p(\theta|\alpha',\beta') = \frac{1}{B(\alpha',\beta')}\theta^{\alpha'-1}(1-\theta)^{\beta'-1}$$
$$\propto \theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}.$$

Posterior and Marginal Likelihood

- Given HTH, we've shown that posterior is

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

- Consider a $\mathcal{B}(\alpha',\beta')$ distribution on θ with $\alpha' = 2+\alpha$ and $\beta' = 1+\beta$,

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

- Probabilities sum to 1: these have same distribution and normalizing constant.

Posterior and Marginal Likelihood

- Given HTH, we've shown that posterior is

$$p(\theta|HTH,\alpha,\beta) = \frac{\frac{1}{B(\alpha,\beta)}\theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}}{p(HTH|\alpha,\beta)} \propto \theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}.$$

- Consider a $\mathcal{B}(\alpha',\beta')$ distribution on θ with $\alpha' = 2+\alpha$ and $\beta' = 1+\beta$,

$$p(\theta|\alpha',\beta') = \frac{1}{B(\alpha',\beta')}\theta^{\alpha'-1}(1-\theta)^{\beta'-1} \propto \theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}.$$

- Probabilities sum to 1: these have same distribution and normalizing constant.
    - Posterior is a beta distribution, $p(\theta|HTH,\alpha,\beta)$ is a $\mathcal{B}(2+\alpha,1+\beta)$ distribution.

Posterior and Marginal Likelihood

- Given HTH, we've shown that posterior is

$$p(\theta|HTH,\alpha,\beta) = \frac{\frac{1}{B(\alpha,\beta)}\theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}}{p(HTH|\alpha,\beta)}$$
$$\propto \theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}.$$

- Consider a $\mathcal{B}(\alpha',\beta')$ distribution on θ with $\alpha' = 2+\alpha$ and $\beta' = 1+\beta$,

$$p(\theta|\alpha',\beta') = \frac{1}{B(\alpha',\beta')}\theta^{\alpha'-1}(1-\theta)^{\beta'-1}$$
$$\propto \theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}.$$

- Probabilities sum to 1: these have same distribution and normalizing constant.
  - Posterior is a beta distribution, $p(\theta|HTH,\alpha,\beta)$ is a $\mathcal{B}(2+\alpha,1+\beta)$ distribution.
  - Marginal likelihood is ratio of posterior and prior normalizing constants,

$$p(HTH|\alpha,\beta) = \frac{B(2+\alpha,1+\beta)}{B(\alpha,\beta)}.$$

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

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 Beta(1,1) prior,
$$\begin{aligned} p(H|HHH) &= \int_0^1 p(H|\theta)p(\theta|HHH)d\theta \\ &= \int_0^1 \mathrm{Ber}(H|\theta)\mathrm{Beta}(\theta|3+\alpha,\beta)d\theta \\ &= \int_0^1 \theta\mathrm{Beta}(\theta|3+\alpha,\beta)d\theta = \mathbb{E}[\theta] \\ &= \frac{(3+\alpha)}{(3+\alpha)+\beta} = \frac{4}{5} = 0.8. \end{aligned}$$

Beta Bernoulli Model Discussion

- If we observe h heads and t tails, the posterior will be $\mathcal{B}(h + \alpha, t + \beta)$.

Beta Bernoulli Model Discussion

- If we observe h heads and t tails, the posterior will be $\mathcal{B}(h+\alpha, t+\beta)$.
  - Posterior summarized by hyper-parameters $\{\alpha, \beta\}$ and counts $\{h, t\}$.

Beta Bernoulli Model Discussion

- If we observe h heads and t tails, the posterior will be $\mathcal{B}(h + \alpha, t + \beta)$.
  - Posterior summarized by hyper-parameters $\{\alpha, \beta\}$ and counts $\{h, t\}$.
- Hyper-parameters α and β are like "pseudo-counts" in our mind before we flip:
  - $\mathcal{B}(1, 1)$ is like seeing one head and one tail before we flip.
    - For HHH, posterior predictive is 0.800.

Beta Bernoulli Model Discussion

- If we observe h heads and t tails, the posterior will be $\mathcal{B}(h+\alpha, t+\beta)$.
  - Posterior summarized by hyper-parameters $\{\alpha, \beta\}$ and counts $\{h, t\}$.
- Hyper-parameters α and β are like "pseudo-counts" in our mind before we flip:
  - $\mathcal{B}(1, 1)$ is like seeing one head and one tail before we flip.
    - For HHH, posterior predictive is 0.800.
  - $\mathcal{B}(3, 3)$ prior is like seeing 3 heads and 3 tails (stronger uniform prior),
    - For HHH, posterior predictive is 0.667.

Beta Bernoulli Model Discussion

- If we observe h heads and t tails, the posterior will be $\mathcal{B}(h+\alpha, t+\beta)$.
  - Posterior summarized by hyper-parameters $\{\alpha, \beta\}$ and counts $\{h, t\}$.
- Hyper-parameters α and β are like "pseudo-counts" in our mind before we flip:
  - $\mathcal{B}(1,1)$ is like seeing one head and one tail before we flip.
    - For HHH, posterior predictive is 0.800.
  - $\mathcal{B}(3,3)$ 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.

Beta Bernoulli Model Discussion

- If we observe h heads and t tails, the posterior will be $\mathcal{B}(h+\alpha, t+\beta)$.
  - Posterior summarized by hyper-parameters $\{\alpha, \beta\}$ and counts $\{h, t\}$.
- Hyper-parameters α and β are like "pseudo-counts" in our mind before we flip:
  - $\mathcal{B}(1,1)$ is like seeing one head and one tail before we flip.
    - For HHH, posterior predictive is 0.800.
  - $\mathcal{B}(3,3)$ 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".

Outline

Bayesian Linear Regression

- In week 2, we argued that L2-regularized linear regression,

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

Bayesian Linear Regression

- In week 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

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

Bayesian Linear Regression

- In week 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 Linear Regression

- In week 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?

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.

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.

Learning the Prior from Data?

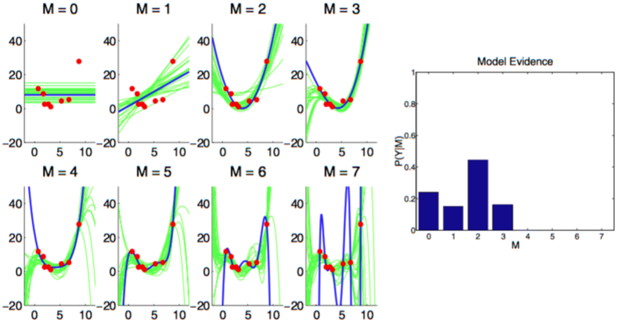
- Can we use the data to set the hyper-parameters?
- In theory: No!
  - It would not be a "prior".
  - It's no longer the right thing to do.
- In practice: Yes!
  - Approach 1: use a validation set or cross-validation as before.
  - Approach 2: optimize the marginal likelihood,

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

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

Type II Maximum Likelihood for Basis Parameter

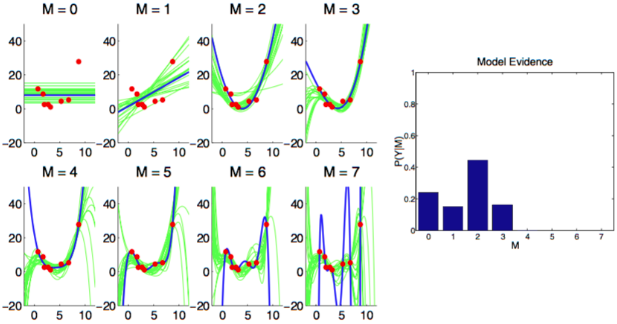
- Consider polynomial basis, and treat degree M as a hyper-parameter:



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

Type II Maximum Likelihood for Basis Parameter

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

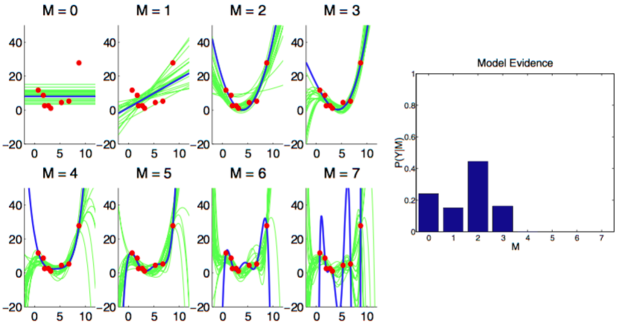


http://www.cs.ubc.ca/~arnaud/stat535/slides5_revised.pdf

- Marginal likelihood (evidence) is highest for $M = 2$.
  - "Bayesian Occam's Razor": prefers simpler models that fit data well.

Type II Maximum Likelihood for Basis Parameter

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

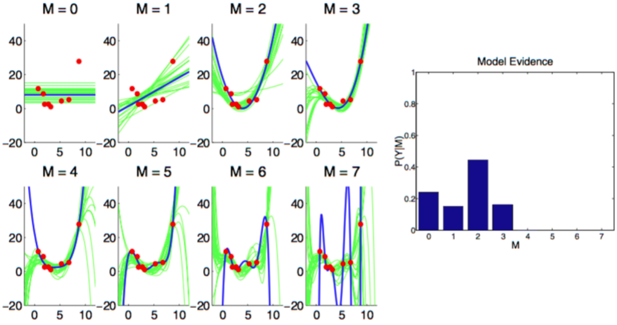


http://www.cs.ubc.ca/~arnaud/stat535/slides5_revised.pdf

- Marginal likelihood (evidence) is highest for $M = 2$.
  - "Bayesian Occam's Razor": prefers simpler models that fit data well.
  - $\sum_{D'} p(D'|m) = 1$, for $M = 7$ we have low $p(D|m)$ since it can fit many datasets.

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.
  - $\sum_{D'} p(D'|m) = 1$, for $M = 7$ we have low $p(D|m)$ since it 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,

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

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

Type II Maximum Likelihood for Regularization Parameter

- Maximum likelihood maximizes probability of data given parameters,

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

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

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

and the integral has closed-form solution because posterior is Gaussian.

Type II Maximum Likelihood for Regularization Parameter

- Maximum likelihood maximizes probability of data given parameters,

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

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

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

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

Type II Maximum Likelihood for Regularization Parameter

- Maximum likelihood maximizes probability of data given parameters,

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

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

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

  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:

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

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 $\lambda = 0$.

Learning Principles

- Maximum likelihood:

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

- MAP:

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

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

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

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,

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

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 maximum likelihood works.
  - You can do gradient descent to optimize the λ_j.

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 maximum likelihood works.
  - You can do gradient descent to optimize the λ_j.
- Weird fact: yields sparse solutions (automatic relevance determination).

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 maximum likelihood works.
  - You can do gradient descent to optimize the λ_j.
- 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 encouages sparsity.

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 maximum likelihood works.
  - You can do gradient descent to optimize the λ_j.
- 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 encouages sparsity.
- Non-convex and theory not well understood, but recent work shows:
  - Never performs worse than L1-regularization, and exists cases where it does better.

Bonus Slide: Overivew of Bayesian Variable Selection

- If we fix λ and use L1-regularization (Bayesian lasso), posterior is not sparse.
  - Probability that a variable is exactly 0 is zero.
  - L1-regularization only lead to sparsity because the MAP point estimate is sparse.
- Type II maximum likelihood leads to sparsity in the posterior because variance goes to zero.
- We can encourage sparsity in Bayesian models using a spike and slab prior:
  - Mixture of Dirac delta function 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"?

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

- 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.

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.
- For dealing with hyper-parameters like λ, we can now do Bayesian inference:
  - Work with posterior over λ, $p(\lambda|X, y, \gamma)$.
  - Computing $p(\lambda_1|X, y, \gamma)/p(\lambda_2|X, y, \gamma)$ is called Bayes factor.

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.
- For dealing with hyper-parameters like λ, we can now do Bayesian inference:
  - Work with posterior over λ, $p(\lambda|X, y, \gamma)$.
  - Computing $p(\lambda_1|X, y, \gamma)/p(\lambda_2|X, y, \gamma)$ is called Bayes factor.
- Bayes factors provide an alternative to classic statistical tests:
  - E.g., we can compute posterior of "fair coin" vs. coin from beta prior.
    - Natural test, but not easy with classic methods.
  - No need for null hypothesis, p-values etc.
  - This month from American Statistical Assocation:
    - "Statement on Statistical Significance and P-Values".
    - `http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108`

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.
- For dealing with hyper-parameters like λ, we can now do Bayesian inference:
  - Work with posterior over λ, $p(\lambda|X, y, \gamma)$.
  - Computing $p(\lambda_1|X, y, \gamma)/p(\lambda_2|X, y, \gamma)$ is called Bayes factor.
- Bayes factors provide an alternative to classic statistical tests:
  - E.g., we can compute posterior of "fair coin" vs. coin from beta prior.
    - Natural test, but not easy with classic methods.
  - No need for null hypothesis, p-values etc.
  - This month from American Statistical Assocation:
    - "Statement on Statistical Significance and P-Values".
    - `http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108`
  - But can only tell you which model is more likely, not whether any model is correct.

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 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 all your assumptions about the world are in the model.
  - Some people try to do this, and have argued that this may be how humans reason.

Discussion of Hierarchical Bayes

- "Super Bayesian" approach:
  - Go up the hierarchy until all your assumptions about the world are in the model.
  - 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, λ, γ, and so on.

Discussion of Hierarchical Bayes

- "Super Bayesian" approach:
  - Go up the hierarchy until all your assumptions about the world are in the model.
  - 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, λ, γ, and so on.
- Key disadvantages:
  - It can be hard to exactly encode your prior beliefs.
  - The integrals get ugly very quickly.

Summary

- Posterior predictive lets us directly model what we want given hyper-parameters.

Summary

- Posterior predictive lets us directly model what we want given hyper-parameters.
- Marginal likelihood is probability seeing data given hyper-parameters.

Summary

- Posterior predictive lets us directly model what we want given hyper-parameters.
- 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.

Summary

- Posterior predictive lets us directly model what we want given hyper-parameters.
- 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.
- Hierarchical Bayes goes even more Bayesian with prior on hyper-parameters.
  - Leads to Bayesian model selection and Bayesian model averaging.

- Next time: can we actually compute these integrals?