CPSC 440: Advanced Machine Learning

Empirical Bayes

Mark Schmidt

University of British Columbia

Winter 2021

Last Time: Bayesian Statistics

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

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

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

- 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:

$$\hat{y}^i \in \underset{\hat{y}}{\operatorname{argmax}}\, p(\hat{y} \mid \hat{x}^i, X, y) \qquad \text{(posterior predictive)}$$

$$\equiv \underset{\hat{y}}{\operatorname{argmax}} \int_w p(\hat{y} \mid \hat{x}^i, w) p(w \mid X, y) dw \qquad \text{(average predictions, weighted by posterior)}$$

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

7 Ingredients of Bayesian Inference (MEMORIZE)

1. Likelihood $p(y \mid X, w)$ (discriminative) or $p(y, X \mid w)$ (generative).
   - Probability of seeing data given parameters.
2. Prior $p(w \mid \lambda)$.
   - Belief that parameters are correct before we've seen data.
3. Posterior $p(w \mid X, y, \lambda)$.
   - Probability that parameters are correct after we've seen data.
   - We won't use the MAP "point estimate", we want the whole distribution.
4. Predictive $p(\tilde{y} \mid \tilde{x}, w)$.
   - Probability of test label $\tilde{y}$ given parameters w and test features $\tilde{x}$.
     - For example, sigmoid function for logistic regression.

7 Ingredients of Bayesian Inference (MEMORIZE)

5. Posterior predictive $p(\tilde{y} \mid \tilde{x}, X, y, \lambda)$.
   - Probability of new data given old, integrating over parameters.
   - This tells us which prediction is most likely given data and prior.

6. Marginal likelihood $p(y \mid X, \lambda)$ (also called "evidence").
   - Probability of seeing data given hyper-parameters (integrating over parameters).
   - We'll use this later for hypothesis testing and setting hyper-parameters.

7. Cost $C(\hat{y} \mid \tilde{y})$.
   - 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.

Review: Decision Theory

- Are we equally concerned about "spam" vs. "not spam".

- 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

- In 340 we discussed predictin $\hat{y}$ given $\hat{w}$ by minimizing expected cost:

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

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

Bayesian Decision Theory

- Bayesian decision theory:
  - Instead of using a MAP estimate $\hat{w}$, we should use posterior predictive,

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

  - 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

Bayesian Linear Regression

- Consider linear regression with Gaussian likelihood and prior,
$$y^i \sim \mathcal{N}(w^T x^i, \sigma^2), \quad w_j \sim \mathcal{N}(0, \lambda^{-1}).$$
- MAP estimation in this model corresponds to L2-regularized linear regression
$$\operatorname*{argmin}_w \frac{1}{2\sigma^2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2.$$
- And the solution is given by a variant on the normal equations:
$$w_{\text{MAP}} = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X^T X + \lambda I\right)^{-1} X^T y.$$
- In 340 we fixed $\sigma^2 = 1$ (changing σ^2 equivalent to changing λ).
  - In the Bayesian framework, both σ^2 and λ affect the predictions.
- To predict on new examples we use $\hat{y} = w_{\text{MAP}}^T \tilde{x}$.

Bayesian Linear Regression

- Consider linear regression with Gaussian likelihood and prior,

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

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

$$w \mid X, y \sim \mathcal{N}\left(w_{\text{MAP}}, \left(\frac{1}{\sigma^2}X^T X + \lambda I\right)^{-1}\right),$$

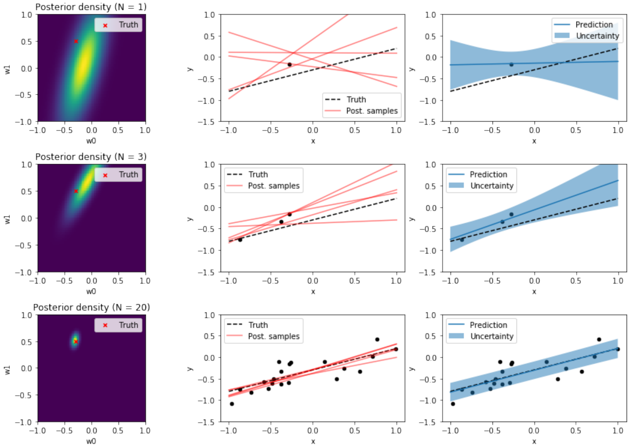
  which is a Gaussian centered at the MAP estimate.
  - The variance tells us how much variance we have around the MAP estimate.
    - Note that usually the MAP is not the mean of the posterior.
- By more tedious Gaussian identities the posterior predictive has the form

$$\tilde{y} \mid X, y, \tilde{x} \sim \mathcal{N}(w_{\text{MAP}}^T \tilde{x}, \sigma^2 + \tilde{x}^T \left(\frac{1}{\sigma^2}X^T X + \lambda I\right)^{-1} \tilde{x}).$$

- Mode of posterior predictive is MAP predictions (not usually the case).
  - And we now have variance of predictions.

Bayesian Linear Regression

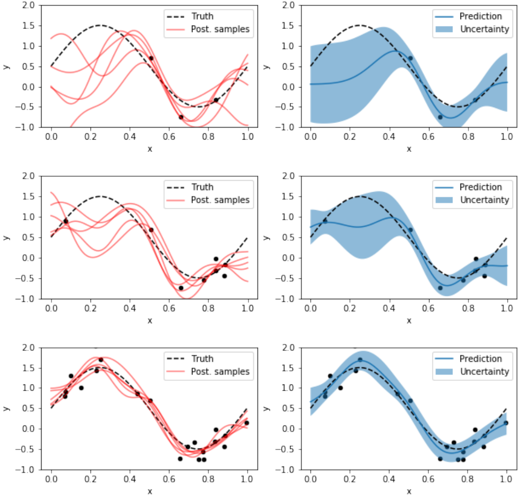
- Bayesian perspective gives us variability in w and predictions:



http://krasserm.github.io/2019/02/23/bayesian-linear-regression

Bayesian Linear Regression

- With Gaussian RBFs are features:



Learning the Prior from Data?

- Can we use the training 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: split into training/validation set or use cross-validation as before.
  - Approach 2: optimize the marginal likelihood ("evidence"):

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

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

Digression: Marginal Likelihood in Gaussian-Gaussian Model

- Suppose we have a Gaussian likelihood and Gaussian prior,

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

- The joint probability of y^i and w_j is the likelihood times the prior:

$$p(y, w \mid X) \propto \exp\left(-\frac{1}{2\sigma^2}\|Xw - y\|^2 - \frac{\lambda}{2}\|w\|^2\right).$$

- The marginal likelihood integrates the joint over the nuisance parameter w,

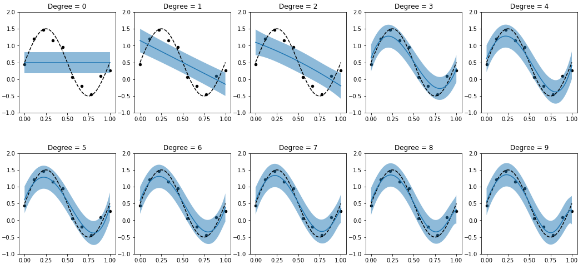
$$p(y \mid X) = \int_w p(y, w \mid X) dw = \int_w p(y \mid X, w) p(w) dw \quad (w \perp X).$$

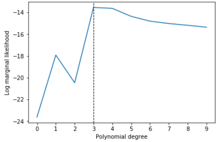
- Solving the Gaussian integral gives a marginal likelihood of

$$p(y \mid X) = \frac{(\lambda)^{d/2}}{(\sigma\sqrt{2\pi})^n \left|\frac{1}{\sigma^2} X^T X + \lambda I\right|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\|Xw_{\mathsf{MAP}} - y\|^2 - \frac{\lambda}{2}\|w^+\|^2\right).$$

Type II Maximum Likelihood for Basis Parameter

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





http://krasserm.github.io/2019/02/23/bayesian-linear-regression

- Marginal likelihood (evidence) is highest for degree 3.
  - "Bayesian Occam's Razor": prefers simpler models that fit data well.
  - $p(y \mid X)$ is smaller for degree 4 polynomials since they can fit more datasets.
  - It's actually non-monotonic it prefers degree 1 and 3 over degree 2.
  - Model selection criteria like BIC are approximations to marginal likelihood as $n \to \infty$.

Type II Maximum Likelihood for Polynomial Basis

- Why is the marginal likelihood higher for degree 3 than 7?
  - Marginal likelihood for degree 3:

$$p(y \mid X) = \int_{w_0} \int_{w_1} \int_{w_2} \int_{w_3} p(y \mid X, w) p(w \mid \lambda) dw$$

  - Marginal likelihood for degree 7:

$$p(y \mid X) = \int_{w_0} \int_{w_1} \int_{w_2} \int_{w_3} \int_{w_4} \int_{w_5} \int_{w_6} \int_{w_7} p(y \mid X, w) p(w \mid \lambda) dw.$$

  - Higher-degree integrates over high-dimensional volume:
    - A non-trivial proportion of degree 3 functions fit the data really well.
    - There are many degree 7 functions that fit the data even better, but they are a much smaller proportion of all degree 7 functions.
  - Warning: this doesn't always work, sometimes becomes degenerate.
    - May need a non-vague prior on the hyper-parameters.

Bayes Factors for Bayesian Hypothesis Testing

- Suppose we want to compare hypotheses:
  - E.g., "this data is best fit with linear model" vs. a degree-2 polynomial.
- Bayes factor is ratio of marginal likelihoods,

$$\frac{p(y \mid X, \text{degree } 2)}{p(y \mid X, \text{degree } 1)}.$$

  - If very large then data is much more consistent with degree 2.
  - A common variation also puts prior on degree.
- A more direct method of hypothesis testing:
  - No need for null hypothesis, "power" of test, p-values, and so on.
  - As usual only says which model is more likely, not whether any are correct.

- American Statistical Assocation:
  - "Statement on Statistical Significance and P-Values".
  - http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108
- "Hack Your Way To Scientific Glory":
  - https://fivethirtyeight.com/features/science-isnt-broken
- "Replicability crisis" in social psychology and many other fields:
  - https://en.wikipedia.org/wiki/Replication_crisis
  - http://www.nature.com/news/big-names-in-statistics-want-to-shake-up-much-maligned-p-value-1.22375
- "T-Tests Aren't Monotonic": https://www.naftaliharris.com/blog/t-test-non-monotonic
- Bayes factors don't solve problems with p-values and multiple testing.
  - But they give an alternative view, are more intuitive, and make assumptions clear.
- Some notes on various issues associated with Bayes factors:
  - http://www.aarondefazio.com/adefazio-bayesfactor-guide.pdf

Type II Maximum Likelihood for Regularization Parameter

- 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 parameters of the prior ("empirical" Bayes).
  - "Optimizing hyper-parameters based on training data".
- Even if we have a complicated model, much less likely to overfit than MLE:
  - Complicated models need to integrate over many more alternative hypotheses.

Learning Principles (MEMORIZE)

- 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:

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

- Bayesian (no "learning"):

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

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

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

Type II Maximum Likelihood for Individual Regularization Parameter

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

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

- 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 may 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.

Summary

- 7 ingredients of Bayesian inference:
  - Likelihood, prior, posterior, predictive, posterior predictive, marginal likelihood, cost.
- Bayesian decision theory:
  - Optimal predictions based on cost functions and rules of probability.
- Marginal likelihood is probability seeing data given hyper-parameters.
  - Bayes factors compute ratios between models to test hypotheses.
- 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.
- Next time: which priors yield closed-form solutions?

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 $w(\lambda) = (X^T X + \lambda I)^{-1} X^T y$.
- 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 can use 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.

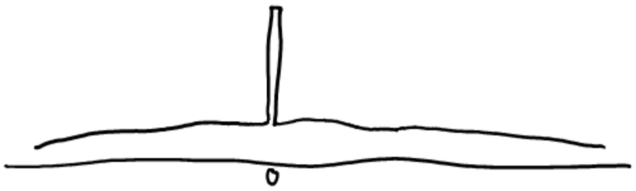
Bayesian Feature Selection

- 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 probability of individual w_j values being 0.

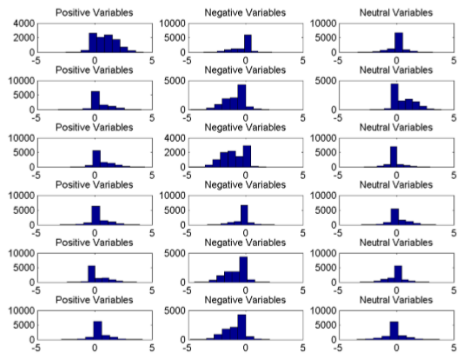
- We can encourage sparsity in Bayesian models using a spike and slab prior:



  - 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.