CPSC 540: Machine Learning

Empirical Bayes

Mark Schmidt

University of British Columbia

Winter 2019

Last Time: Bayesian Statistics

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

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

- 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:

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

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

Coin Flipping Example: MAP Approach

- MAP vs. Bayesian for a simple coin flipping scenario:
  1. Our likelihood is a Bernoulli,
  $$p(H \mid \theta) = \theta.$$
  2. Our prior assumes that we are in one of two scenarios:
     - The coin has a 50% chance of being fair ($\theta = 0.5$).
     - The coin has a 50% chance of being rigged ($\theta = 1$).
  3. Our data consists of three consecutive heads: 'HHH'.

- What is the probability that the next toss is a head?
  - MAP estimate is $\hat{\theta} = 1$, since $p(\theta = 1 \mid HHH) > p(\theta = 0.5 \mid HHH)$.
  - So MAP says the probability is 1.

  - But MAP overfits: we believed there was a 50% chance the coin is fair.

Coin Flipping Example: Posterior Distribution

- Bayesian method needs posterior probability over θ,

$$p(\theta = 1 \mid HHH) = \frac{p(HHH \mid \theta = 1)p(\theta = 1)}{p(HHH)} \quad \text{(Bayes rule)}$$

$$\text{(marg and prod rule)} = \frac{p(HHH \mid \theta = 1)p(\theta = 1)}{p(HHH \mid \theta = 0.5)p(\theta = 0.5) + p(HHH \mid \theta = 1)p(\theta = 1)}$$

$$= \frac{(1)(0.5)}{(1/8)(0.5) + (1)(0.5)} = \frac{8}{9},$$

and similarly we have $p(\theta = 0.5 \mid HHH) = \frac{1}{9}$.

- So given the data, we should believe with probability $\frac{8}{9}$ that coin is rigged.
  - There is still a $\frac{1}{9}$ probability that it is fair that MAP is ignoring.

Coin Flipping Example: Posterior Predictive

- Posterior predictive gives probability of head given data and prior,

$$
\begin{aligned}
&p(H \mid HHH) = p(H, \theta = 1 \mid HHH) + p(H, \theta = 0.5 \mid HHH) \\
&= p(H \mid \theta = 1, HHH) p(\theta = 1 \mid HHH) \\
&+ p(H \mid \theta = 0.5, HHH) p(\theta = 0.5 \mid HHH) \\
&= (1)(8/9) + (0.5)(1/9) = 0.94.
\end{aligned}
$$

- So the correct probability given our assumptions/data is 0.94, and not 1.
  - Though with a different prior we would get a different answer.
- Notice that there was no optimization of the parameter θ:
  - In Bayesian stats we condition on data and integrate over unknowns.
- In Bayesian stats/ML: "all parameters are nuissance parameters".

Coin Flipping Example: Discussion

Comments on coin flipping example:

- Bayesian prediction uses that HHH could come from fair coin.
- As we see more heads, posterior converges to 1.
  - MLE/MAP/Bayes usually agree as data size increases.
- If we ever see a tail, posterior of $\theta = 1$ becomes 0.

- If the prior is correct, then Bayesian estimate is optimal:
  - Bayesian decision theory gives optimal action incorporating costs.
- If the prior is incorrect, Bayesian estimate may be worse.
  - This is where people get uncomfortable about "subjective" priors.

- But MLE/MAP are also based on "subjective" assumptions.

Bayesian Model Averaging

- In 340 we saw that model averaging can improve performance.
  - E.g., random forests average over random trees that overfit.
- But should all models get equal weight?
  - What if we find a random decision stump that fits the data perfectly?
    - Should this get the same weight as deep random trees that likely overfit?
  - In science, research may be fraudulent or not based on evidence.
    - Should "vaccines cause autism" or climate change denial models get equal weight?
- In these cases, naive averaging may do worse.

Bayesian Model Averaging

- Suppose we have a set of m probabilistic classifiers w_j
  - Previously our ensemble method gave all models equal weights,

$$p(\tilde{y} \mid \tilde{x}) = \frac{1}{m}p(\tilde{y} \mid \tilde{x}, w_1) + \frac{1}{m}p(\tilde{y} \mid \tilde{x}, w_2) + \cdots + \frac{1}{m}p(\tilde{y} \mid \tilde{x}, w_m).$$

- Bayesian model averaging (following rules of probability) weights by posterior,

$$p(\tilde{y} \mid \tilde{x}) = p(w_1 \mid X, y)p(\tilde{y} \mid \tilde{x}, w_1) + p(w_2 \mid X, y)(\tilde{y} \mid \hat{x}, w_2) + \\ \cdots + p(w_m \mid X, y)p(\tilde{y} \mid \tilde{x}, w_m).$$

- So we should weight by probability that w_j is the correct model.
  - Equal weights assume all models are equally probable and fit data equally well.

Bayesian Model Averaging

- Weights are posterior, so proportional to likelihood times prior:

$$p(w_j \mid X, y) \propto \underbrace{p(y \mid X, w_j)}_{\text{likelihood}} \underbrace{p(w_j)}_{\text{prior}}.$$

- Likelihood gives more weight to models that predict y well.
- Prior should gives less weight to models that are likely to overfit.

- This is how rules of probability say we should weight models.
  - It's annoying that it requires a "prior" belief over models.
  - But as $n \to \infty$, all weight goes to "correct" model[s] w^* as long as $p(w^*) > 0$.

Bayes for Density Estimation and Generative/Discriminative

- We can use Bayesian approach for density estimation:
  - With data D and parameters θ we have:
    1. Likelihood $p(D \mid \theta)$.
    2. Prior $p(\theta)$.
    3. Posterior $p(\theta \mid D)$.

- We can use Bayesian approach for supervised learning:
  - Generative approach (naive Bayes, GDA) are density estimation on X and y:
    1. Likelihood $p(y, X \mid w)$.
    2. Prior $p(w)$.
    3. Posterior $p(w \mid X, y)$.

  - Discriminative approach (logistic regression, neural nets) just conditions on X:
    1. Likelihood $p(y \mid X, w)$.
    2. Prior $p(w)$.
    3. Posterior $p(w \mid X, y)$.

7 Ingredients of Bayesian Inference

1. Likelihood $p(y \mid X, w)$.
    - 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

4. 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.
5. 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.
6. 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:

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

- 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

- We know that L2-regularized linear regression,

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

corresponds to MAP estimation in the model

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

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

$$w \mid X, y \sim \mathcal{N}\left(\frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X^TX + \lambda I\right)^{-1} X^Ty, \left(\frac{1}{\sigma^2}X^TX + \lambda I\right)^{-1}\right).$$

- 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 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 given by

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

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

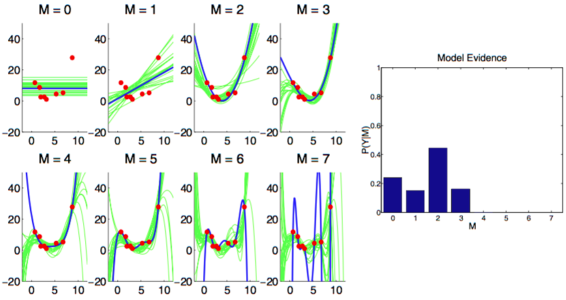
$$p(y \mid X, \lambda) = \int_w p(y, w \mid X, \lambda) dw.$$

- Solving the Gaussian integral gives a marginal likelihood of

$$p(y \mid X, \lambda) \propto |C|^{-1/2} \exp\left(-\frac{y^T C^{-1} y}{2}\right), \quad C = \sigma^2 I + \frac{1}{\lambda} X X^T.$$

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.
  - $p(y \mid X, \lambda)$ is small for $M = 7$, since 7-degree polynomials can fit many datasets.
  - It's actually non-monotonic in M: it prefers $M = 0$ and $M = 2$ over $M = 1$.
  - 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 high for degree 2 but not degree 7?
  - Marginal likelihood for degree 2:

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

  - Marginal likelihood for degree 7:

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

  - Higher-degree integrates over high-dimensional volume:
    - A non-trivial proportion of degree 2 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.

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 can only tell you which model is 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

Learning Principles

- Maximum likelihood:

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

- MAP:

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

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

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

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

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

Type II Maximum Likelihood for 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 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.

Type II Maximum Likelihood for Individual Regularization Parameter

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

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

- Too expensive for cross-validation, but type II MLE works.
  - You can do gradient descent to optimize the λ_j.
- Weird fact: this yields sparse solutions.
  - "Automatic relevance determination" (ARD)
  - Can send $\lambda_j \to \infty$, concentrating posterior for w_j at exactly 0.
    - It tries to "remove some of the integrals".
  - This is L2-regularization, but empirical Bayes naturally encourages sparsity.
- Non-convex and theory not well understood:
  - Tends to yield much sparser solutions than L1-regularization.

Type II Maximum Likelihood for Other Hyper-Parameters

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

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

- You can also use type II MLE to optimize these values.

- The "automatic relevance determination" selects training examples ($\sigma_i \rightarrow \infty$).
  - This is like the support vectors in SVMs, but tends to be much more sparse.

- Type II MLE can also be used to learn kernel parameters like RBF variance.
  - Do gradient descent on the σ values in the Gaussian kernel.

- It will also do something sensible if you use it to choose number of clusters k.
  - Or number of states in hidden Markov model, number of latent factors in PCA, etc.

- Bonus slides: Bayesian feature selection gives probability that w_j is non-zero.
  - Posterior is much more informative than standard sparse MAP methods.

Summary

- Bayesian model averaging and decision theory:
  - Model averaging and decision theory based on rules of probability.
- Marginal likelihood is probability seeing data given hyper-parameters.
- 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: putting a prior on the prior and relaxing IID

Gradient on Validation/Cross-Validation Error

- It's also possible to do gradient descent on λ to optimize validation/cross-validation error of model fit on the training data.
- For L2-regularized least squares, define $w(\lambda) = (X^TX + \lambda I)^{-1}X^Ty$.
- You can use chain rule to get derivative of validation error E_{valid} with respect to λ:

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

- For more complicated models, you 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.

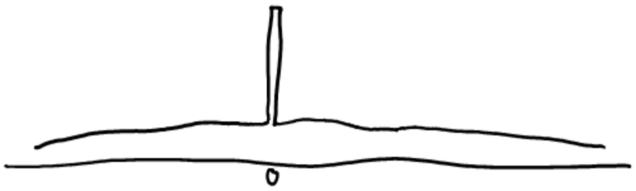
Bayesian Feature Selection

- 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 probabiliy of being 0.

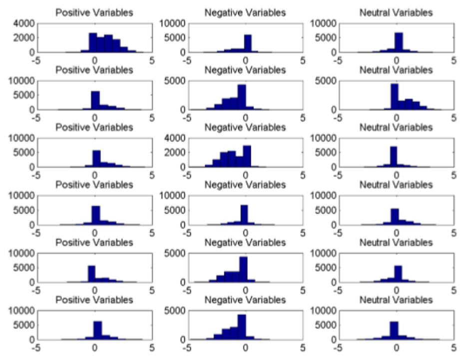
- We can encourage sparsity in Bayesian models using a spike and slab prior:



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

Bayesian Feature Selection

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



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