Empirical Bayes

CPSC 440/550: Advanced Machine Learning

cs.ubc.ca/~dsuth/440/24w2

University of British Columbia, on unceded Musqueam land

2024-25 Winter Term 2 (Jan–Apr 2025)

Last time: Multivariate Gaussians

- Fitting multivariate Gaussians:
  - MLE is again sample mean / covariance
  - Conjugate prior for the mean with known covariance: Gaussian
  - Non-conjugate MAP estimate for the covariance: $\hat{\mathbf{\Sigma}} + \lambda \mathbf{I}$
  - Conjugate prior exists (normal-Wishart)
- Generative classifiers with Gaussians: LDA, QDA
- Bayesian linear regression
  - Basic form: same probabilistic model where ridge regression is the MAP
  - Bayesian learning gives a posterior distribution over $w \mid \mathbf{X}, \mathbf{y}$
  - and a corresponding posterior predictive distribution for $\tilde{y} \mid \tilde{x}, \mathbf{X}, \mathbf{y}$

Outline

Setting hyperparameters

- Bayesian linear regression has hyperparameters σ^2, λ
  - If choosing feature transform / kernel function, potentially many more

- The usual validation set approach to choosing them:
  - Split into a training and validation set
  - For each hyperparameter value (in a grid, selected randomly, ...):
    - Compute some measure of test error, e.g. negative log-likelihood
  - Choose the hyperparameter setting with the lowest error

- Advantage: directly tries to achieve good performance on new data
- Disadvantages:
  - Can easily overfit to the validation set if model is flexible enough
  - Slow; many possible hyperparameter settings to try

Learning the prior from data?

- An alternative approach to fitting hyperparameters: empirical Bayes
- Maximizes the training likelihood given the hyperparameters

$$\hat{\alpha} \in \arg\max_{\alpha} p(\mathbf{X} \mid \alpha) = \arg\max_{\alpha} \int p(\mathbf{X} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \alpha)\, \mathrm{d}\boldsymbol{\theta}$$

  - Note: α could be any number of hyperparameters, $\boldsymbol{\theta}$ any number of parameters
- $p(\mathbf{X} \mid \alpha)$ is called the "marginal likelihood" or "evidence"
- It's the denominator when we do MAP: $p(\boldsymbol{\theta} \mid \mathbf{X}, \alpha) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)}{p(\mathbf{X}|\alpha)}$
- Can think of as MLE for the hyper-parameters
  - Empirical Bayes also called "type II maximum likelihood" or "evidence maximization"

- Advantages:
  - Often fast! Sometimes closed-form, sometimes gradient descent (if conjugate prior)
  - Doesn't require a separate validation set
- Disadvantages:
  - It doesn't look at the fit on new data, just on training data
  - Can overfit the marginal likelihood

Marginal likelihood with conjugate priors

- Marginal likelihood has a nice closed form when using conjugate priors
- When $x \mid \theta \sim \mathrm{Bern}(\theta)$, $\theta \sim \mathrm{Beta}(\alpha, \beta)$, let $B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta$:

$$
\begin{aligned}
p(\mathbf{X} \mid \alpha, \beta) &= \int p(\mathbf{X} \mid \theta)\, p(\theta \mid \alpha, \beta)\, \mathrm{d}\theta \\
&= \int \theta^{n_1}(1-\theta)^{n_0}\, \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \mathrm{d}\theta \\
&= \frac{1}{B(\alpha,\beta)} \int \theta^{(n_1+\alpha)-1}(1-\theta)^{(n_0+\beta)-1} \mathrm{d}\theta \\
&= \frac{B(n_1+\alpha, n_0+\beta)}{B(\alpha,\beta)}
\end{aligned}
$$

- This result is generally true up to a multiplicative constant for conjugate priors

Learning principles

- Maximum likelihood:

$$\hat{\theta} \in \arg\max_{\theta} p(\mathbf{X} \mid \theta) \qquad \text{use } p(\tilde{x} \mid \hat{\theta})$$

- Maximum a posteriori (MAP):

$$\hat{\theta} \in \arg\max_{\theta} p(\theta \mid \mathbf{X}, \alpha) \qquad \text{use } p(\tilde{x} \mid \hat{\theta})$$

- Bayesian with fixed prior:

$$\text{use } p(\tilde{x} \mid \mathbf{X}, \alpha) = \int p(\boldsymbol{\theta} \mid \mathbf{X}, \alpha) p(\tilde{x} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

- Empirical Bayes:

$$\hat{\alpha} \in \arg\max_{\alpha} p(\mathbf{X} \mid \alpha); \qquad \text{use } p(\tilde{x} \mid \mathbf{X}, \hat{\alpha}) = \int p(\boldsymbol{\theta} \mid \mathbf{X}, \hat{\alpha}) p(\tilde{x} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

Bayesian hierarchy

- MLE can do weird things
  - Can give zero probability for events not in training
    - "I flipped a coin twice and it was heads both times, it must *always* be heads"
  - Generally, might pick highly "unlikely" model that exactly fits training data
- MAP helps by adding a prior, but still commits to one parameter
- Bayesian inference makes optimal decisions if your likelihood/prior are "correct"
  - No "optimization bias" because there's no optimization
  - Predictions exactly follow rules of probability
  - Only works if the model (prior + likelihood) is good
- Empirical Bayes uses data to find a good prior
  - Tends to be less sensitive to overfitting than normal MLE
  - Can still overfit; it's just MLE in a "less sensitive" model!

Bayesian hierarchy

- Empirical Bayes can overfit in its choice of the hyper-parameter α
- So, maybe we should put a hyper-prior on α (with hyper-hyper-parameters)
- But we're still uncertain about the choice of α,
  so really maybe we should marginalize over all possible choices of α
  - Can do Bayesian inference over parameters and hyper-parameters together
  - Helps avoid overfitting
  - Usually don't have a convenient "conjugate hyper-prior" to work with

- This process depends on having a good hyper-prior
- Maybe we should fit it from data by maximizing the marginal likelihood...
- And maybe we should use a hyper-hyper-prior to make a good choice...
- In practice, model *tends* to be less sensitive at each level, so don't need to go forever

Wikipedia: "Turtles all the way down"

Outline

1. Empirical Bayes (in general)
2. Empirical Bayes for Bayesian linear regression

Setting Hyper-Parameters with Empirical Bayes

- To set hyper-parameters like σ^2 and λ, we could use a validation set
  - (Can do efficient leave-one-out cross-validation at least for ridge regression)

- But could also use empirical Bayes and optimize the marginal likelihood,

$$\hat{\sigma}^2, \hat{\lambda} \in \underset{\sigma^2, \lambda}{\arg\max}\, p(\mathbf{y} \mid \mathbf{X}, \sigma^2, \lambda)$$

- The marginal likelihood integrates over the parameters w,

$$p(\mathbf{y} \mid \mathbf{X}, \sigma^2, \lambda) = \int_w p(\mathbf{y}, w \mid \mathbf{X}, \sigma^2, \lambda) \mathrm{d}w = \int_w p(\mathbf{y} \mid \mathbf{X}, w, \sigma^2) p(w \mid \lambda) \mathrm{d}w \quad (w \perp\!\!\!\perp X)$$

- This is the marginal in a product of Gaussians, which is (with some work):

$$p(\mathbf{y} \mid \mathbf{X}, \sigma^2, \lambda) = \frac{(\lambda)^{d/2} (\sigma\sqrt{2\pi})^{-n}}{\sqrt{\det\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{X}w_{\mathsf{MAP}} - \mathbf{y}\|^2 - \frac{\lambda}{2}\|w_{\mathsf{MAP}}\|^2\right)$$

  - You could run gradient descent on the negative log of this to set hyper-parameters
    - You could do "projected" gradient or reparameterize to handle constraints

Setting Hyper-Parameters with Empirical Bayes

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

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

- Too expensive for cross-validation, but can still do empirical Bayes
  - You can do projected gradient descent to optimize the λ_j

- Weird fact: this yields sparse solutions
  - It can send some $\lambda_j \to \infty$, concentrating posterior for w_j at exactly 0
  - This is L2-regularization, but empirical Bayes naturally encourages sparsity
    - "Automatic relevance determination" (ARD)

- Non-convex, theory not really well understood
- Tends to yield much sparser solutions than L1 regularization

Setting Hyper-Parameters with Empirical Bayes

- Consider also having a hyper-parameter $\sigma^{(i)}$ for each i,

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

- You can also use empirical Bayes to optimize these hyper-parameters

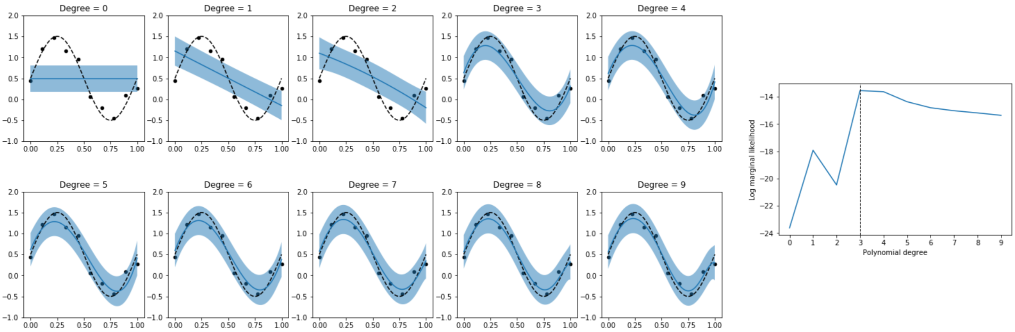
- The "automatic relevance determination" selects training examples ($\sigma_i \to \infty$)
  - This is like the support vectors in SVMs, but tends to be much more sparse

- Empirical Bayes can also be used to learn kernel parameters like RBF variance
  - Do gradient descent on the lengthscales in the Gaussian kernel

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

Choosing Polynomial Degree with Empirical Bayes

- Using empirical Bayes to choose degree hyper-parameter with polynomial basis:



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
  - $p(y \mid X, \sigma^2, \lambda, k)$ is smaller for degree 4 polynomials since they can fit more datasets
  - It's non-monotonic: it prefers degree 1 and 3 over degree 2
  - Model selection criteria like BIC approximate marginal likelihood as $n \to \infty$

Choosing Polynomial Degree with Empirical Bayes

- Why is the marginal likelihood higher for degree 3 than 7?
- Marginal likelihood for degree 3 (ignoring conditioning on hyper-parameters):

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

- Marginal likelihood for degree 7:

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

- 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

Choosing Between Bases with Empirical Bayes

- We could compare marginal likelihood between different non-linear transforms:

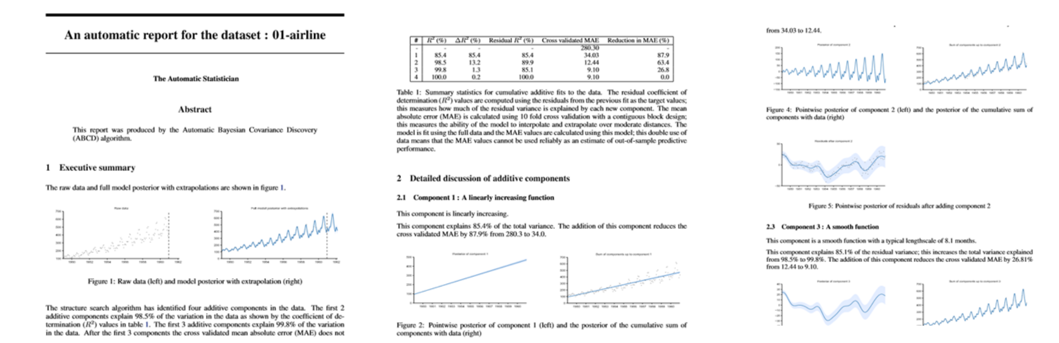
$$p(\mathbf{y} \mid \mathbf{X}, \text{polynomial basis}) > p(\mathbf{y} \mid \mathbf{X}, \text{Gaussian RBF as basis})?$$

- This is the idea behind Bayes factors for hypothesis testing (see bonus slides)
  - Alternative to classic hypothesis tests like t-tests
- Usual warning: empirical Bayes can sometimes become degenerate
  - May need a non-vague prior on the hyper-parameters
- But we could have a hyper-prior over possible non-linear transformations
  - Use empirical Bayes in this hierarchical model to learn basis and parameters

Application: Automatic Statistician

bonus!

- Can be viewed as an automatic statistician:
  `http://www.automaticstatistician.com/examples`



An automatic report for the dataset : 01-airline

The Automatic Statistician

Abstract

This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified four additive components in the data. The first 2 additive components explain 98.5% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 3 additive components explain 99.8% of the variation in the data. After the first 3 components the cross validated mean absolute error (MAE) does not

#	R^2 (%)	ΔR^2 (%)	Residual R^2 (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	280.30	-
1	85.4	85.4	85.4	34.03	87.9
2	98.5	13.2	89.9	12.44	63.4
3	99.8	1.3	85.1	9.10	26.8
4	100.0	0.2	100.0	9.10	0.0

Table 1: Summary statistics for cumulative additive fits to the data. The residual coefficient of determination (R^2) values are computed using the residuals from the previous fit as the target values; this measures how much of the residual variance is explained by each new component. The mean absolute error (MAE) is calculated using 10 fold cross validation with a contiguous block design; this measures the ability of the model to interpolate and extrapolate over moderate distances. The model is fit using the full data and the MAE values are calculated using this model; this double use of data means that the MAE values cannot be used reliably as an estimate of out-of-sample predictive performance.

2 Detailed discussion of additive components

2.1 Component 1 : A linearly increasing function

This component is linearly increasing.

This component explains 85.4% of the total variance. The addition of this component reduces the cross validated MAE by 87.9% from 280.3 to 34.0.

Figure 2: Pointwise posterior of component 1 (left) and the posterior of the cumulative sum of components with data (right)

from 34.03 to 12.44.

Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Figure 5: Pointwise posterior of residuals after adding component 2

2.3 Component 3 : A smooth function

This component is a smooth function with a typical lengthscale of 8.1 months.

This component explains 85.1% of the residual variance; this increases the total variance explained from 98.5% to 99.8%. The addition of this component reduces the cross validated MAE by 26.81% from 12.44 to 9.10.

Summary

- Empirical Bayes for linear regression
  - Can use marginal likelihood to noise variance(s) and regularization parameters(s)
  - Can also select which non-linear transforms to use
    - Bayesian Occam's razor: can encourage sparsity and simplicity
- Bayesian logistic regression
  - Gaussian prior is not conjugate so need approximations

- Next time: how to approximate for non-conjugate priors

Gradient of Validation/Cross-Validation Error

bonus!

- 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

Bayesian Feature Selection

bonus!

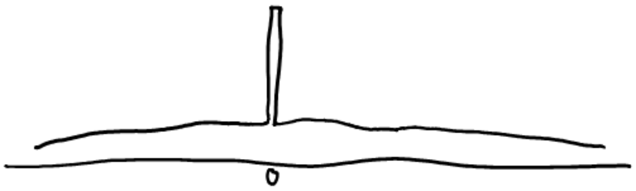
- Classic feature selection methods don't work when $d >> n$:
  - AIC, BIC, Mallow's, adjusted-R^2, and L1-regularization return very different results.
- Here maybe all we can hope for is posterior probability of $w_j = 0$.
  - Consider all models, and weight by posterior the ones where $w_j = 0$.
- 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

bonus!

- 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:

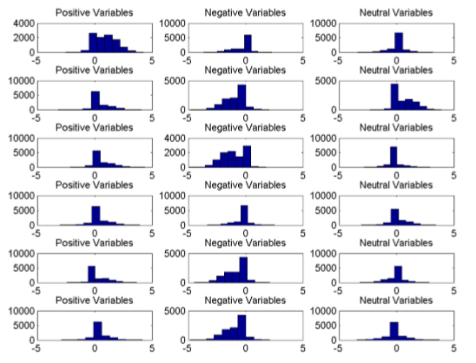


  - Mixture of Dirac delta function at 0 and another prior with non-zero variance.
  - Places non-zero posterior weight at exactly 0.
  - Posterior is still non-sparse, but answers the question:
    - "What is the probability that variable is non-zero"?

Bayesian Feature Selection

bonus!

- 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.

Bayes Factors for Bayesian Hypothesis Testing

bonus!

- 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.

bonus!

- 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