Online Bounds for Bayesian Algorithms

Sham Kakade
University of Pennsylvania

Joint work with Andrew Ng

Methodologies

- Bayesian:
  - often make strong assumptions (in the prior) on the data generation process
  - optimality guaranteed here
- Online Learning:
  - adversarial setting (against Nature) where there is no data generation process
  - weaker notion of optimality

Motivation

How do Bayesian algorithms fare in a more adversarial setting?

- Often Bayesian methods make assumptions we don't believe (eg i.i.d. assumptions)
- Often models chosen for computational tractability
- Bayes rule looks like an 'expert' algorithm so we would expect it to perform well.

Outline

- The Setup
- A General Online Bound
- Bounds for Bayesian Model Averaging
- Bounds for Maximum Aposteriori estimation

The Setup

The Setting

- inputs x in R^n and outputs y in R
- sequence of examples $S = \{(x_1, y_1), \ldots, (x_T, y_T)\}$
  - not specifying generative model for S
  - S_t is the subsequence from time 1 to t
- Using a model,
  - at time t, we predict with $p(y|x_t, S_{t-1})$

The Model

- Consider a generalized linear model:

$$p(y|x,\theta) = p(y|\theta^T x)$$

- For example,
  - linear least squares: $p(y|x,\theta) \sim \mathcal{N}(\theta^T x, \sigma^2)$
  - logistic regression: $p(y|x,\theta) \sim \sigma(\theta^T x)^y (1 - \sigma(\theta^T x))^{1-y}$
- Assume a prior: $p(\theta) \sim \mathcal{N}(\vec{0}, \nu^2 I_n)$

Loss at a Timestep

- at time $t\text{-}1$, we have a posterior $-\log p(\theta|S_{t-1})$
- Bayesian Model Averaging:
- $$p(y|x_t, S_{t-1}) = \int_\theta p(y|x_t,\theta)p(\theta|S_{t-1})d\theta$$
- at time t, our loss is $-\log p(y_t|x_t, S_{t-1})$

Total Losses

- Our loss:

$$L_{BMA}(S) = \sum_{t=1}^{T} -\log p(y_t|x_t, S_{t-1})$$

- "Expert" loss:

$$L_\theta(S) = \sum_{t=1}^{T} -\log p(y_t|x_t, \theta)$$

- Another loss w.r.t. Q:

$$L_Q(S) = \int_\theta Q(\theta) L_\theta(S) d\theta$$

A Useful Bound

A General Online Bound

- **Theorem:** For all sequences S and distributions Q:

$$L_{BMA}(S) \leq L_Q(S) + KL(Q||p)$$

- **Proof:**
  - similar to Freund & Schapire
  - show that:

$$L_{BMA}(S) = L_Q(S) + KL(Q||p) - KL(Q||p(\theta|S_T))$$

Bounds for BMA

An Upper Bound

- Suppose $|\partial^2 \log p(y|\theta^T x)/(\partial\theta^T x)^2| \leq c$
  - for linear regression $c = 1/\sigma^2$
  - for logistic regression $c=1$
- Theorem: Then
  - $L_{BMA}(S) \leq L_\theta(S) + \frac{1}{2\nu^2}||\theta||^2 + \frac{n}{2}\log\left(1 + \frac{Tc\nu^2}{n}\right)$
  - the second term is a penalty from our prior
  - the log term is how fast the loss grows

Proof Idea

- Recall

$$L_{BMA}(S) \leq L_Q(S) + KL(Q||p)$$

- For Q, choose $N(\theta, \varepsilon^2 I_n)$
- Then use derivative bound to show $L_Q(S)$ is close to $L_\theta(S)$
- Optimize ε

A Lower Bound

- Theorem: For linear regression, the upper bound is tight.
- Proof: exhibit a "worst case" sequence
  - We can restrict Nature to use a generative model for S that is i.i.d.
  - Nature uses a $p(y|x_t)$ that is in our model
  - in this sense, the worst case isn't much different than an average case

Bounds for MAP

MAP Estimation

- Use the max $\hat{\theta}_{t-1}$ of $p(\theta|S_{t-1})$ for the prediction
  - the loss is $-log p(y_t|x_t, \hat{\theta}_{t-1})$
  - recall BMA has loss $-log p(y_t|x_t, S_{t-1})$
- In practice, MAP often used (computational reasons?)
- We consider both cases of linear and logistic regression.

Ridge Regression

- Use the squared loss:

$$L_{MAP}(S) = \frac{1}{2}\sum_{t=1}^{T}(y_t + \hat{\theta}_{t-1}^{T}x_t)^2$$

  - which is essentially just the sum log loss

- Corollary: The MAP loss is a multiplicative factor of $\nu^2 + \sigma^2$ worse.

- Vovk has a better bound for this case

  - the algorithm is related to ridge regression (but it is nonlinear)

Why is MAP worse?

- **Theorem (Lower Bound):** The upper bound for MAP cannot have a multiplicative factor of 1.

- Compare

  - BMA's loss

  - $$L_{BMA}(S) = \sum_{t=1}^{T} \frac{1}{2s_t^2}(y_t - \hat{\theta}_{t-1}^T x_t)^2 + \log\sqrt{2\pi s_t^2}$$

  - to MAP's loss

$$L_{MAP}(S) = \frac{1}{2}\sum_{t=1}^{T}(y_t + \hat{\theta}_{t-1}^T x_t)^2$$

MAP for logistic regression

- BMA is intractable
- MAP is widely used
  - essentially regularized logistic regression
  - involves solving a convex program
- Theorem: The loss for MAP is multiplicatively worse by a factor of 4.

Conclusions

- Some Bayesian algorithms perform well in an adversarial setting
- Open Problem: Can the dimensionality dependence on the bounds be removed with further assumptions?