Statistical Techniques in Robotics (16-831, F10) Lecture #16 (Thursday October $14^{th}$)

# Bayesian Linear Regression (BLR) Part 1

*Lecturer: Drew Bagnell* *Scribe: Heather Knight, Xinjilefu* [1]

## 1 Bayes' Online Learning with a Prior

### 1.1 Standard Weighted Majority with a Prior

- Numer of experts: $N$
- Prior: $p_i$ ($p_i \geq 0$ and $\sum_i p_i \geq 1$)
- Set initial non-uniform weights to each expert: $w_i^1 = Np_i$
- Each expert makes prediction $y_i \in \{0, 1\}$
- **for** $t = 1, \cdots, T$

Predict:

- Predict 1 **if**

$$\sum_{y_i=1} w_i^t \geq \sum_{y_i=0} w_i^t \tag{1}$$

- **else**, Predict 0

Update:

- If expert $e_i$ made a mistake, $w_i^{t+1} \leftarrow \frac{1}{2} w_i^t$

Analysis of Algorithm:

- Total weights of the experts will decrease over time with mistakes $W = \sum_i w_i$
- Weight of the best expert $w^* \leq W$
- $m$ is the total number of mistakes predicted by the algorithm
- $m^*$ are the number of mistakes made by the best expert:

$$w^* = 2^{-m^*} N p^* \tag{2}$$

[1] Some content adapted from previous scribes: Kevin Lipkin, Alvaro Collet-Romea, Laura Lindzey and Hans Pirnay

- The total weight $W$ is at most

$$W \leq N \left(\frac{4}{3}\right)^{-m} \tag{3}$$

- Thus, since $w^* \leq W$

$$2^{-m^*} N p^* \leq N \left(\frac{4}{3}\right)^{-m} \tag{4}$$

$$-m^* + \log_2 p^* \leq -mc \tag{5}$$

  where $c = \log_2(\frac{4}{3})$

- Therefore, the total mistakes made by the algorithm are bounded by:

$$m \leq \frac{m^* + \log_2(\frac{1}{p^*})}{c} \tag{6}$$

Weighted majority using prior:

- No dependence on the number of expert $N$
- Because of prior, infinite number of experts are possible (except for the weight update)
- If you see "log n" where n is some discrete set of experts, think hidden uniform distribution
- Every learning algorithm has a prior, the prior for the Weighted Majority is all experts are equally good ($p_i = \frac{1}{N}$)
- Priors in hypothesis space correspond to weights on experts

### 1.2 General Weighted Majority Update

- Bayes' Rule is a special case of weighted majority
- Predict:
  - Choose expert $e_i$ in proportion to $\frac{w_i}{\sum_j w_j}$
  - Predict the same as what expert $e_i$ predicts
- Receive Loss: $l_t(i)$
- Update Weights:
  - $w_i^{t+1} = w_i^t e^{-\alpha l_t(i)}$
    or, use first term of Taylor Series expansion:
  - $w_i^{t+1} = w_i^t(1 - \alpha l_t(i))$
    shows that Bayes Rule has no regret properties
- Expert $e_i$'s prediction is a probability distribution: $p_i(y)$

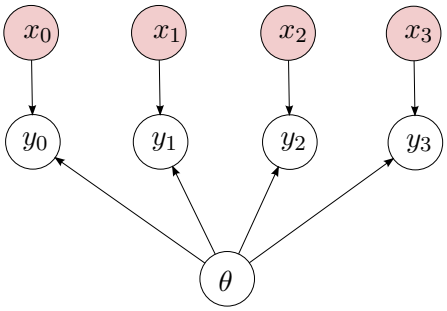


Figure 1: Graphic model of BLR

## 2 Bayes' Linear Regression

- $\theta$ = weight vector
- $x_t$ = set of features at every timestep
- $y_t$ = true prediction of outcome

We want to derive how to update our representation given a series of data $D = \{x, y\}$ up to timestep $t$ and $x_{t+1}$, the probability of an observation $\tilde{y}_{t+1}$ is

$$p(\tilde{y}_{t+1}|x_{t+1}, D) = \int p(\tilde{y}_{t+1}|x_{t+1}, D, \theta) \cdot p(\theta|D) d\theta \tag{7}$$

$$= \int p(\tilde{y}_{t+1}|x_{t+1}, \theta) \cdot p(\theta|D) d\theta \tag{8}$$

In BLR, the prior of the weight vector $\theta$ is a Gaussian where $\theta \sim N(\mu_0, \Sigma_0)$.

$$p(\theta) = \frac{1}{z} \exp\{-(\theta - \mu_0)^T \Sigma_0^{-1} (\theta - \mu_0)\} \tag{9}$$

This is called the Moment Parameterization of a Guassian, where $\Sigma_0$ is positive-definite and symmetric.

The Natural Parameterization is given by:

$$p(\theta) = \frac{1}{z} \exp\{J_0^T \theta - \frac{1}{2} \theta^T P_0 \theta\} \tag{10}$$

where

$$P_0 = \Sigma_0^{-1} \tag{11}$$

$$J_0 = P_0 \mu_0 \tag{12}$$

We can calculate $p(\theta|D)$ recursively using the Bayes' Theorem

$$p(\theta|D) = \frac{1}{z} p(y_t|\theta, D) p(\theta|D) \tag{13}$$

where we know the likelihood $p(y_t|D,\theta) = p(y_t|x_t,\theta)$ is given by a Gaussian $\mathcal{N}(\theta^T x_t, \sigma_t^2)$

$$p(y_t|x_t,\theta) = \frac{1}{z}\exp\{\frac{-(\theta^T x_t - y_t)^2}{2\sigma^2}\} \tag{14}$$

The updating rule for $p(\theta|D)$ at timestep $t$ is given by multiplication of two exponential functions. Adding the exponent of the prior to that of the likelihood yields

$$-\frac{1}{2\sigma^2}\left(\theta^T x_t - y_t\right)^2 + J_{t-1}^T\theta - \frac{1}{2}\theta^T P_{t-1}\theta \tag{15}$$

collecting terms to find updates $J_t$ and $P_t$:

$$= -\frac{1}{2\sigma^2}\left(\theta^T x_t x_t^T\theta - 2\theta^T x_t y_t + y_t^2\right) + J_{t-1}^T\theta - \frac{1}{2}\theta^T P_{t-1}\theta \tag{16}$$

$$= \left(\frac{x_t^T y_t}{\sigma^2} + J_{t-1}^T\right)\theta - \frac{1}{2}\theta^T\left(\frac{x_t x_t^T}{\sigma^2} + P_{t-1}\right)\theta - \frac{y_t^2}{2\sigma^2} \tag{17}$$

Since this all happens in the exponent of an exponential function, the constant $y_t^2/\sigma^2$-term can be shifted into the regularizing constant $z$. Thus, the update rules for $J_t$ and $P_t$ are

$$J_t = \frac{y_t x_t}{\sigma^2} + J_{t-1} \tag{18}$$

$$P_t = \frac{x_t x_t^T}{\sigma^2} + P_{t-1} \tag{19}$$

1. In a gaussian model, a new datapoint always lowers the variance - this downgrading of the variance does not always make sense
2. If you believe there are outliers, this model won't work for you
3. The variance is not a function of $y_t$. The precision is only affected by input not output. This is a consequence of having the same $\sigma$ (observation error) everywhere in space.