Statistical Techniques in Robotics (16-831, F08) Lecture #13 (Tuesday October 7nd)

Bayes' Linear Regression Part 2

Lecturer: Drew Bagnell *Scribe: Kevin Lipkin*

1 Bayes' Online Learning with Prior

- p_i = prior
- Set initial weights to each expert: $w_i = Np_i$
- Each expert makes prediction y_i
- Predict:
  - Predict 1 If:
$$\sum_{y=1} w_i \geq \sum_{y=0} w_i \tag{1}$$
  - Else, Predict 0
- Update:
  - If expert e_i made a mistake, $w_i = 1/2w_i$
- Analysis of Algorithm:
  - Total weights of the experts W = $\sum_i \mathrm{w}_i$
  - 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:
$$w^* = 2^{-m^*} Np^* \tag{2}$$
$$W \leq N(\frac{4}{3})^{-M} \tag{3}$$
  - Thus, since $w^* \leq W$
$$2^{-m^*} Np^* \leq N(\frac{4}{3})^{-M} \tag{4}$$
$$-m^* + \log p^* \leq -Mc \tag{5}$$
    Where $c = \log_2(4/3)$
  - Therefore, the total mistakes made by the algorithm are bounded by:
$$M \leq \frac{m^* + \log(\frac{1}{p^*})}{c} \tag{6}$$

- Weighted majority using prior thus has:
  - No dependence on N
  - Because of prior, infinite sets of experts are possible
  - If you see "log n" where n is some discrete set of experts, think hidden uniform distribution
  - Every learning algorithm has a prior - some are more explicit than others
  - Priors in hypothesis space correspond to weights on experts

2 General Weighted Majority Update

- Bayes' Rule is a special case of weighted majority
- Predict:
  - Choose expert 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:
  - $w_i = w_i e^{-\alpha l_t(i)}$
    or, use first term of Taylor Series expansion:
  - $w_i = w_i(1 - \alpha l_t(i))$
- Expert i's prediction is a probability distribution: $p_i(y)$
- Standard loss for making a probabilistic prediction is log-loss:

$$l_t(i) = \log\left(p_i(y_t)\right) \tag{7}$$

  Where y_t is the true observation

- Plugging the log-loss into the weight update rule:

$$w_i = w_i e^{-\alpha \log\left(p_i(y_t)\right)} \tag{8}$$

- This simplifies to:

$$w_i = w_i [p_i(y_t)]^\alpha \tag{9}$$

- Which, when $\alpha = 1$, is Bayes' Rule exactly. According to Bayes' Rule: $p(i|y) = p(i)p(y|i)$. In this case, $p(i|y)$ is equivalent to w_{t+1}, $p(i)$ is equivalent to w_t, and $p(i|y)$ is $p_i(y_t)$.

3 Bayes' Linear Regression

- θ = Weight Vector
- x_t = set of features
- y_t = outcome
- Use Gaussian distribution for likelihood term:

$$p(y|x,\theta) = \frac{1}{z} e^{\frac{-(\theta^T x - y)^2}{2\sigma^2}} \tag{10}$$

  This is called the Moment Parameterization of a Guassian.

- Prior term is a multidemensional guassian:

$$p(\theta) = \frac{1}{z} e^{-(\theta-\mu)^T \Sigma^{-1} (\theta-\mu)} \tag{11}$$

  Where Σ is positive-definite and symmetric

- The Natural Parameterization of (11) is:

$$p(\theta) = \frac{1}{z} e^{J^T\theta - \frac{1}{2}\theta^T P \theta} \tag{12}$$

- $p(\theta|y,x) = p(\theta)p(y|x,\theta)$ = (12)*(10):

$$p(\theta)p(y|x,\theta) = \frac{1}{z} e^{\frac{-(\theta^T x - y)^2}{2\sigma^2} + J^T\theta - \frac{1}{2}\theta^T P \theta} \tag{13}$$

- Combining like terms leaves us with a form very similar to our prior expression (12). Thus, we can update the values of J and P:

$$J' = J + \frac{yx^T}{\sigma^2} \tag{14}$$

$$P' = P + \frac{xx^T}{\sigma^2} \tag{15}$$