Statistical Techniques in Robotics (16-831, F09) Lecture #15 (Tuesday, October 13, 2009)

Online Learning with Prior, Bayes Linear Regression (Part 1)

Lecturer: Drew Bagnell *Scribe: Alvaro Collet-Romea*

1 Bayes' Online Learning with Prior on experts

- Assume a set of N experts
- Set initial weights to each expert:
  $w_i = Np_i$, where p_i is a prior on experts ($p_i \geqslant 0$ and $\sum_i p_i = 1$)
- 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 $W = \sum_i w_i$
  - 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
- After t iterations, the weight w_i of an expert with m_i mistakes is given by:
$$w_i = 2^{-m_i} Np_i \tag{2}$$
- And the global weight must be at most:
$$W \leq N \left(\frac{4}{3}\right)^{-M} \tag{3}$$
- Thus, since $w_i \leq w^* \leq W$
$$2^{-m_i} Np_i \leq N \left(\frac{4}{3}\right)^{-M} \tag{4}$$
$$-m_i + \log p_i \leq -MC \tag{5}$$
  Where $C = \log_2 \left(\frac{4}{3}\right)$

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

$$M \leq \frac{m_i + \log(\frac{1}{p_i})}{C} \tag{6}$$

- From Eq. 6, we can see that m_i is a linear term and the rest is constant.
- Some observations on weighted majority using prior:
  - 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
  - $\log \frac{1}{p_i}$ is the code length, under an optimal code, for hypothesis i.
  - *learning* is very correlated with *compression*: we can represent a sequence as our best hypothesis plus the number of mistakes we make with it.
  - Example of prior: $p_i = \frac{1}{N}$

2 Bayes Linear Regression

2.1 Parameterizations of a Gaussian

Moment parameterization:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad p(X) = \frac{1}{\mathcal{Z}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right) \tag{7}$$

Natural parameterization:

$$X \sim \mathcal{N}(J, P) \quad \Rightarrow \quad p(X) = \frac{1}{\mathcal{Z}} \exp\left(J^T X - \frac{1}{2} X^T P X\right) \tag{8}$$

Conversion from Natural to Moment parameterization:

$$J = \Sigma^{-1}\mu \quad P = \Sigma^{-1} \tag{9}$$

2.2 Update rule in Bayes LR

- θ = Weight Vector
- x_t = set of features
- $y_t \sim \mathcal{N}\left(\theta^T x, \sigma^2\right)$ = prediction

- Use Gaussian distribution for likelihood term:

$$p(y_t|x_t, \theta) = \frac{1}{\mathcal{Z}} \exp\left(\frac{-\left(\theta^T x_t - y_t\right)^2}{2\sigma^2}\right) \tag{10}$$

- Prior term is a N-dimensional Gaussian:

$$\theta \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad p(\theta) = \frac{1}{\mathcal{Z}} \exp\left(-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1} (\theta - \mu)\right) \tag{11}$$

  Where Σ is a covariance matrix, positive-definite and symmetric

- The Natural Parameterization, equivalent to Eq. 11, is:

$$p(\theta) = \frac{1}{\mathcal{Z}} \exp\left(J^T \theta - \frac{1}{2}\theta^T P \theta\right) \tag{12}$$

- $p(\theta|y, x_t) \sim p(\theta)p(y_t|x_t, \theta) =$ (Eq. 12)*(Eq. 10)

$$p(\theta)p(y_t|x_t, \theta) = \frac{1}{\mathcal{Z}} \exp\left(\frac{-(\theta^T x_t - y_t)^2}{2\sigma^2} + J^T \theta - \frac{1}{2}\theta^T P \theta\right) \tag{13}$$

- Combining linear and quadratic terms we can reach an expression similar to Eq. 11, and matching them we obtain the update rules for J and P:

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

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

Observations:

- $P = \Sigma^{-1}$ implies that P will grow the more certain you are about your predictions
- If we never see a feature x^i of x_t, P won't change in that direction, i.e. there will be no update for x^i.