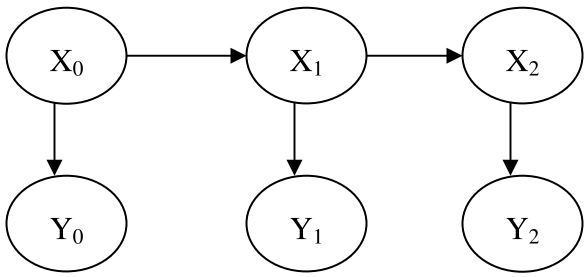
Statistical Techniques in Robotics (16-831, F08) Lecture #27 (11/25/2008)

Conditional Random Field and Filters

Lecturer: Drew Bagnell Scribe: Saurabh Taneja

Conditional Models



P(y|x) is complicated Bayes independence assumption

$p(\vec{x}, \vec{y}) = p(\vec{y} \mid \vec{x}).p(\vec{x})$ Generative Description

Conditional (discriminative) description

$p(\vec{x}, \vec{y} \mid \lambda) = p(\vec{y} \mid \vec{x}, \lambda).p(\vec{x})$

Principle of maximum entropy

$x = \{1,2,...,6\}$ Dice case

$$\arg \underset{p \in P}{Max}\, H(p) = \sum_i p(x_i) \log \frac{1}{p(x_i)} \tag{1}$$

where p is the probability distribution over all the set P of probability distributions

Average Properties

E[x] = 3.5 (additional constraint)

so for a given piece of data the task is to find the distribution that maximizes (1)

Another Constraint

$E[(x-\mu)^2] = 1$

This problem originated from Statistical Physics where for e.g. Physicists had to find out the distribution of velocity of molecules of a gas and they estimated it from the pressure of the gas.

Method of Lagrange Multipliers

It is one of the ways of achieving our goal of finding an appropriate probability distribution.

It penalizes the difference between objective function and constraint

$$\max_{p\in P} -\sum_i p(x_i)\log\frac{1}{p(x_i)} - \lambda(E[x_i]-3.5) - \mu(\sum_i p(x_i)-1)$$

Convex Fn　　　　　　Linear constraint

Hence there is only one global minimum.

$$\frac{\partial L}{\partial \lambda} = 0, \frac{\partial L}{\partial \mu} = 0, \frac{\partial L}{\partial p_i} = 0$$

$$\frac{\partial L}{\partial p_i} = 0 \Rightarrow p_i = \frac{1}{z}\exp(-\lambda x_i)$$

where z is a normalizer

so if our linear constraint is　　　　E[f(x)] = a

then　　　$p_i = \frac{1}{z}\exp\{-\lambda f(x_i)\}$

and for multiple constraints　　$p_i = \frac{1}{z}\exp\{-\lambda_1 f(x_i) - \lambda_2 g(x_i) - \lambda_3 h(x_i)\}$

For a given mean and variance　　$p_i = \frac{1}{z}\exp\{-\lambda_1 x_i - \lambda_2 \mu_i^2\}$　　which is a Gaussian.

Hence, if we have a mean and variance then Gaussian is the distribution that makes least assumptions.

For finding out the value of λ we need to solve an optimization problem using gradient descent (or something else).

Gradient Descent

Suppose p(x|y) is a simple classification problem of whether x is a rock or a bush.

So now we have to $\max E_{p(y)}[H\{p(x \mid y)\}]$

features f_1, f_2, f_3

therefore $$p(x \mid y, \lambda) = \frac{1}{z}\exp\{-\lambda_1 f_1 - \lambda_2 f_2 - \lambda_3 f_3\}$$

calculate $$\arg \underset{\lambda}{Max} \log[\prod_i p(x_i \mid y_i, \lambda).p(\lambda_i)]$$

we need to solve $$\frac{\partial}{\partial \lambda}[\sum_i \log(x_i \mid y_i, \lambda) + \log p(\lambda_i)]$$

which is $$\frac{\partial}{\partial \lambda}[\sum_i \log \frac{\exp(-\lambda^T F)}{z} + \log p(\lambda_i)]$$

now $$\frac{\partial}{\partial \lambda}[\Sigma \log \frac{\exp(-\lambda^T F)}{z}] = \frac{\partial}{\partial \lambda}[\Sigma\{\log \exp(-\lambda^T F) - \log z(\lambda)\}]$$

$$\frac{\partial}{\partial \lambda}\log z(\lambda) = \frac{1}{z}\frac{\partial z}{\partial \lambda}$$

Therefore $\frac{1}{z}\frac{\partial z}{\partial \lambda} = -E_{p(x|y)} f$ i.e. expectation of f under the distribution

Hence the gradient rule states

$$\lambda_i += \alpha.\sum_i E_{p(x|y)}[f_i] - f_i$$