Probabilistic Artificial Intelligence

Problem Set 4

Nov 4, 2019

Please check out the Moodle webpage for exam-like questions on the following link: moodle-app2.let.ethz.ch/course/view.php?id=11902

1. Particle filter

Suppose that you have a robot, which is moving randomly through an 1-dimensional environment. You want to track the robot's position, x, which is discretized to integer values, $x \in \mathbb{Z}$. The robot's movement is modeled as a random walk,

$$x_{t+1} = x_t + \epsilon_t, \tag{1}$$

where ϵ_t is uniformly distributed and can take integer values in $[-3, 3]$. To track the robot, a sensor that measures the distance to the robot has been placed at the origin. The measurement model is

$$y_t = (x_t + \eta_t)^2, \tag{2}$$

where η_t is distributed according to

$$P(\eta_t) = \begin{cases} 0.6 & \text{if } \eta_t = 0 \\ 0.2 & \text{if } |\eta_t| = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

You want to use a particle filter with six particles to track the robot's position. At initial time, the robot is at the origin, $x_0 = 0$. Hence, the particles are initialized to $x_i = 0$, $i \in \{0, 1, 2, 3, 4, 5\}$.

(i) You draw samples from the distribution of ϵ_0 and obtain $(-1, -1, 0, 1, 2, 3)$. What is the position of the particles after the prediction update?

(ii) You obtain a measurement, $y_1 = 1$. What are the weights of the individual particles?

(iii) Are five particles enough to accurately estimate the state? Why/Why not?

(iv) Why would a Kalman filter not work reliably in this case?

2. Bayesian networks and Markov chains

Consider the query $P(R|S = t, W = t)$ in the following Bayesian network, and how Gibbs

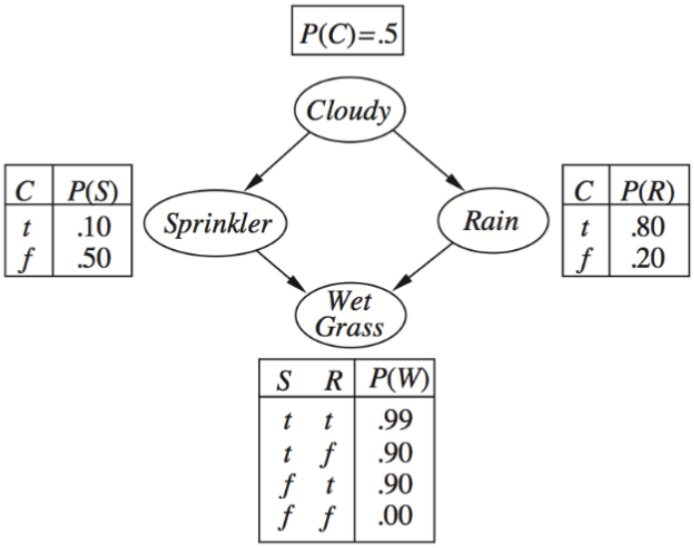

Figure 1: Bayesian Network

sampling can answer it.

(i) How many states does the Markov chain have?

(ii) Calculate the transition matrix T containing $P(X_{t+1} = y \mid X_t = x)$ for all x, y.

(iii) What does T^2, the square of the transition matrix, represent?

(iv) What about T^n as $n \to \infty$?

(v) Explain how to do probabilistic inference in Bayesian networks, assuming that T^n is available. Is this a practical way to do inference?

3. Assumed Density Filters

Let $p(x)$ be any distribution on $\mathbf{R}^n$. Let $q(x)$ be a multivariate normal distribution with mean μ and covariance Σ, i.e.:

$$q(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu))$$

Prove that the forward KL divergence between $p(x)$ and $q(x)$, $D(p||q)$ is minimzed by moment matching, i.e.: $\mu = \mathbf{E}_{x\sim p(\cdot)}[x]$ and $\Sigma = \mathbf{E}_{x\sim p(\cdot)}\left[(x-\mu)(x-\mu)^\top\right]$

Hint: Use formulas 57, 61, and 86 of the Matrix Cookbook http://www.math.uwaterloo.ca/~hwolkowi//matrixcookbook.pdf

4. Assumed Density Filters for Object Tracking in the Presence of Clutter

We assume an object is located at a position $\boldsymbol{\theta} \in \mathbb{R}^d$. At time step t for $t = 1, \ldots, n$ we are able to observe the noisy position of this object $\mathbf{x}_t \in \mathbb{R}^d$ through the model given below:

$$p(x_t|\theta) = (1-w)\mathcal{N}(x_t; \theta, I_d) + w\mathcal{N}(x_t; 0, 10I_d) \tag{4}$$

Where $\mathcal{N}(.;\mu,\Sigma)$ denotes the standard d-dimensional, normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and $0 < w < 1$ is a known mixing parameter. We let $\boldsymbol{\theta} \in \mathbb{R}^d$ have a prior distribution:

$$p(\theta) = \mathcal{N}(\theta; 0, I_d)$$

Therefore the joint distribution of $\boldsymbol{\theta}$ and n independent observations $\boldsymbol{D} \triangleq \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$ is given by:

$$p(\theta, D) = p(\theta)\prod_{t=1}^{n} p(x_t|\theta)$$

The goal here is to approximate the true posterior $p(\theta|x_1,\ldots,x_t)$ at time step $1 \le t \le n$. As a new data point x_{t+1} is received at time step $t+1$ we would like to update our posterior to accommodate for the new data point.

We assume the posterior has an (approximate) distribution that is spherical Gaussian parametrized by its mean and variance. More precisely at time step $t \in [n]$,we approximate the true posterior by:

$$q^{(t)}(\theta) = \mathcal{N}(\theta; \mu^{(t)}, \sigma^{(t)2} I_d)$$

Where $\mu^{(t)} \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}^+$. We set $q^{(0)}(t) = p(\theta)$ (equal to the prior).

(i) Draw a Bayesian network corresponding to this model. What are the hidden and observed random variables at time t?

(ii) For $t = 1, \ldots, n$ and assuming that $q^{(t-1)}(\theta) = p(\theta|x_1, \ldots, x_{t-1})$ i.e. $q^{(t-1)}(\theta)$ is exactly equal to the posterior at time $(t-1)$, show that the posterior at time t is given by:

$$\hat{p}(\theta) \triangleq p(x_1,\ldots,x_t|\theta) = \frac{q^{(t-1)}(\theta)p(x_t|\theta)}{\int q^{(t-1)}(\theta)p(x_t|\theta)}$$

(iii) In this step we approximate the true posterior $\hat{p}(\theta)$ with $q^{(t)}(\theta)$ as follows:

$$\mu^{(t)}, \sigma^{(t)} = \operatorname*{argmin}_{\mu^t,\sigma^t} KL(\hat{p}(\theta)||q^{(t)}(\theta))$$

Show that:

$$\mu^{(t)} = \mathbb{E}_{\boldsymbol{\theta}\sim q^{(t)}}[\boldsymbol{\theta}] = \mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}] \tag{5}$$

$$\sigma^{(t)} d + \mu^{(t)\top} \mu^{(t)} = \mathbb{E}_{\boldsymbol{\theta} \sim q^{(t)}}[\boldsymbol{\theta}^\top \boldsymbol{\theta}] = \mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}^\top \boldsymbol{\theta}] \tag{6}$$

Hint: Take derivatives of the KL with respect to $\mu^{(t)}$ and $\sigma^{(t)}$ and set them to zero as in question 4

(iv) In this step we show how one can compute $\mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}]$ and $\mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}^\top \boldsymbol{\theta}]$. Let us define:

$$Z(\mu^{(t)}, \sigma^{(t)}) = \int p(x_i|\theta) q^{(t)}(\theta) d\theta$$

By differentiating $\log Z$ w.r.t $\mu^{(t)}$ and $\sigma^{(t)}$ show that (we drop the dependence of $\mu^{(t)}$ and $\sigma^{(t)}$ on (t) for notational convenience) :

$$\mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}] = \mu + \sigma \nabla_\mu \log(Z(\mu, \sigma)) \tag{7}$$

$$\mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}^\top \boldsymbol{\theta}] - \mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}]^\top \mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}] = \sigma d - \sigma^2 (\nabla_\mu^\top \nabla_\mu - 2\nabla_\sigma) \log Z(\mu, \sigma) \tag{8}$$

Where $\nabla_\mu^\top \nabla_\mu \log Z = \|\nabla_\mu \log Z\|^2$

(v) By inserting $p(x_i|\theta)$ from equation (4) derive update equations for $\mu^{(t)}$ and $\sigma^{(t)}$ given $\mu^{(t-1)}$ and $\sigma^{(t-1)}$ and x_t.