Machine Learning
Exercise 7

Marc Toussaint

Machine Learning & Robotics lab, U Stuttgart

Universitätsstraße 38, 70569 Stuttgart, Germany

June 28, 2014

1 Sum of 3 dices

You have 3 dices (potentially fake dices where each one has a different probability table over the 6 values). You're given all three probability tables $P(D_1)$, $P(D_2)$, and $P(D_3)$. Write down the equations and an algorithm (in pseudo code) that computes the conditional probability $P(S|D_1)$ of the sum of all three dices conditioned on the value of the first dice.

2 Product of Gaussians

Slide 05:34 defines a 1D and n-dimensional Gaussian distribution. See also Wikipedia, if necessary. Multiplying probability distributions is a fundamental operation, and multiplying two Gaussians is needed in many models. From the definition of a n-dimensional Gaussian, prove the general rule

$$\mathcal{N}(x \,|\, a, A)\, \mathcal{N}(x \,|\, b, B) \;\propto\; \mathcal{N}(x \,|\, (A^{-1} + B^{-1})^{-1}(A^{-1}a + B^{-1}b), (A^{-1} + B^{-1})^{-1}) \;,$$

where the proportionality $\propto$ allows you to drop all terms independent of x.

Note: The so-called canonical form of a Gaussian is defined as $\mathcal{N}[x \,|\, \bar a, \bar A] = \mathcal{N}(x \,|\, \bar A^{-1} \bar a, \bar A^{1})$; in this convention the product reads much nicher: $\mathcal{N}[x \,|\, \bar a, \bar A]\, \mathcal{N}[x \,|\, \bar b, \bar B] \propto \mathcal{N}[x \,|\, \bar a + \bar b, \bar A + \bar B]$. You can first prove this before proving the above, if you like.

3 Gaussian Processes

Consider a Gaussian Process prior $P(f)$ over functions defined by the mean function $\mu(x) = 0$, the γ-exponential covariance function

$$k(x, x') = \exp\{-|(x - x')/l|^\gamma\}$$

and an observation noise $\sigma = 0.1$. We assume $x \in \mathbb{R}$ is 1-dimensional. First consider the standard squared exponential kernel with $\gamma = 2$ and $l = 0.2$.

a) Assume we have two data points $(-0.5, 0.3)$ and $(0.5, -0.1)$. Display the posterior $P(f|D)$. For this, compute the mean posterior function $\widehat f(x)$ and the standard deviation function $\widehat\sigma(x)$ (on the 100 grid points) exactly as on slide 06:10, using $\lambda = \sigma^2$. Then plot $\widehat f$, $\widehat f + \widehat\sigma$ and $\widehat f - \widehat s$ to display the posterior mean and standard deviation.

b) Now display the posterior $P(y^*|x^*, D)$. This is only a tiny difference from the above (see slide 06:8). The mean is the same, but the variance of y^* includes additionally the obseration noise σ^2.

c) Repeat a) & b) for a kernel with $\gamma = 1$.