LINKÖPING UNIVERSITY
Dept. of Computer and Information Science
Division of Statistics and Machine Learning
Jose M. Peña and Mattias Villani

2016-10-20
Advanced Machine Learning, 6 hp
732A96

Computer Exam - Advanced Machine Learning (732A96), 6 hp

Time:	2-6 PM
Allowable material:	- Paper copy of the book 'Pattern Recognition and Machine Learning' by Bishop. - 100 page PDF file set up by the student (readable at full scale) available through the computer exam system. - Slides from the lectures, available through the computer exam system.
Teachers:	Jose M. Peña. Phone: $070-0895280$ (available from 2 PM) Mattias Villani. Phone: $070-0895205$ (available from 3.15 PM)
Grades:	Maximum number of credits on the exam: 20. Maximum number of credits on each exam question: 5. A=19-20 points B=17-18 points C=12-16 points D=10-11 points E=8-9 points F=0-7 points The total exam score is rounded to the nearest integer.

Full credit requires clear and well motivated answers.

1. Graphical Models

(a) Learn a Bayesian network (both structure and parameters) from the Asia dataset that is distributed in the bnlearn package. Use any learning algorithm from the bnlearn package and settings that you consider appropriate. Use the Bayesian network learned to compute the conditional probability of person having visited Asia given that the person has bronchitis and the X-rays came positive, i.e. $p(A|X = TRUE, B = TRUE)$. Use both the approximate and exact methods. (2 p)

(b) There are 29281 DAGs with five nodes. Compute approximately the fraction of the 29281 DAGs that represent an independence model that can be represented with a Markov network. You may want to use the function `skeleton` of the bnlearn package, which outputs the undirected graph that results from dropping the directions of the edges in the input graph. (2 p)

(c) Explain how to perform probabilistic reasoning (i.e. compute a conditional probability distribution) in a Bayesian network. Please, be as detailed as possible but do not use more than 250 words. (1 p)

2. Hidden Markov Models

Recall Lab 2 where you were asked to build a HMM for modeling the behavior of a robot that walks around a ring. The ring is divided into 10 sectors. At any given time point, the robot is in one of the sectors and decides with equal probability to stay in that sector or move to the next sector. You do not have direct observation of the robot. However, the robot is equipped with a tracking device that you can access. The device is not very accurate though: If the robot is in the sector i, then the device will report that the robot is in the sectors $[i-2, i+2]$ with equal probability.

(a) You are asked to extend the HMM built in Lab 2 as follows. The observed random variable has now 11 states, corresponding to the 10 sectors of the ring plus a 11th state to represent that the tracking device is malfunctioning. If the robot is in the sector i, then the device will report that it is malfunctioning with probability 0.5 and that the robot is in the sectors $[i-2, i+2]$ with probability 0.1 each. Implement the extension just described by using the HMM package. Moreover, consider the observations $1, 11, 11, 11$, i.e. the tracking device reports sector 1 first, and then malfunctioning for three time steps. Compute the most probable path using the smoothed distribution and the Viterbi algorithm. Explain why the paths differ, if they do. (2 p)

(b) You are asked to modify the HMM built in Lab 2 as follows. The ring has now only five sectors. If the robot is in the sector i, then the tracking device will report that the robot is in the sectors $[i-1, i+1]$ with equal probability. The rest of the sectors receive zero probability. The robot now spends at least two time steps in each sector. You are asked to implement this modification. In particular, the regime's minimum duration should be implemented implicitly by duplicating hidden states and the observation model, i.e. do not use increasing or decreasing counting variables. (2 p)

(c) Explain how to learn the parameters of a HMM given a sample of observations, i.e. no hidden states are included in the sample. Please, be as detailed as possible but do not use more than 250 words. (1 p)

3. Gaussian Processes

The file KernelCode.txt distributed with the exam contains code to construct the following three kernel functions in the kernlab format:

$$k_1(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{r^2}{2\ell^2}\right)$$
$$k_2(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha}$$
$$k_3(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right).$$

where $r = |\mathbf{x} - \mathbf{x}'|$.

(a) Now, let $\sigma_f = \ell = 1$ and assume a prior mean of zero for the process. Compute and plot the covariance function corresponding to the three different kernel functions k_1, k_2, k_3. Use three different values of α: 1/2, 2 and 20 in k_2. Plot all covariance functions in the same graph over the domain `r<-seq(0,4,by=0.01)`. Argue about the dependence properties of the Gaussian process from these different covariance functions. For example, what do you expect simulated realizations from the different kernel function to look like? What does this graph tell you about the effect of the hyperparameter α on the process? (1.5 p)
[Hint 1: the function `kernelMatrix` in kernlab may be useful, but is not strictly necessary to solve this exercise. Hint 2: you may find it useful to simulate realizations of the process for the different kernels. This is not required to get full score, but may help you when arguing about the properties of the covariance functions].

(b) The file GPdata.RData contains two variables `y` and `x`. Load the variables into memory with the R command `load("GPdata.RData")`. Compute the posterior distribution of f in the model

$$y = f(x) + \varepsilon, \quad \varepsilon \sim N(0, 0.5^2).$$

Use both the squared exponential (k_1) kernel and the k_3 kernel with $\ell = 1$ and $\sigma_f = 1$. Your answer should be in the form of a scatter plot of the data overlayed with curves for

i. the posterior mean of f
ii. 95% probability intervals for f
iii. 95% prediction intervals for a new data point y

Interpret the results under i)-iii), and explain the difference between the results from ii) and iii). Discuss the differences in results from using the two kernels. Would you prefer one kernel over the other?

Use the `gausspr` function in the kernlab package for i), but not for ii) and iii).

[Hint: $Cov(f) = K(X_\star, X_\star) - K(X_\star, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1} K(X, X_\star)$ and remember that `%*%` does matrix multiplication and `solve` computes inverses in R] (2 p)

(c) (No need to do any computations here). Discuss how a Bayesian would handle the case where the kernel hyperparameters are unknown. Discuss how Bayesian analysis can be used to choose between the kernels k_1, k_2 and k_3. (1.5 p)

4. State-space models

The data set radiation_data.Rda contains $T = 365$ daily measurements of radioactivity (unit: micro sieverts per hour) between December 5, 2012 and December 4, 2013 from the Township village center near the Fukushima power plant. Assume the following local linear trend model for the data

$$\begin{aligned} y_t &= \alpha_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2 = 0.16) \\ \alpha_t &= \alpha_{t-1} + \beta_{t-1} + v_t^{(1)}, \quad v_t^{(1)} \sim N\left(0, \sigma_{v^{(1)}}^2 = 0.035\right) \\ \beta_t &= \beta_{t-1} + v_t^{(2)}, \quad v_t^{(2)} \sim N\left(0, \sigma_{v^{(2)}}^2 = 3.06 \cdot 10^{-12}\right), \end{aligned}$$

where ϵ_t, $v_t^{(1)}$ and $v_t^{(2)}$ are independent. Assume a priori that α_0 and β_0 are independent with priors $p(\alpha_0) = N(10, 10^2)$ and $p(\beta_0) = N(0, 10^2)$.
You should use the R package `dlm` to solve this problem. Load the data frame by the command `load("Radiation_data.Rda")`. The variable of interest (y_t) is `dose` which you can access by `Radiation_data$dose`. Only make plots of the results when asked to (so you don't waste your valuable time!).

(a) Let the state vector be $x_t = (\alpha_t, \beta_t)^T$ and formulate the model as a dynamic linear Gaussian state-space model. That is, write

$$\begin{aligned} y_t &= Fx_t + \epsilon_t, \quad \epsilon_t \sim N(0, V) \\ x_t &= Gx_{t-1} + v_t, \quad v_t \sim N(0, W) \\ p(x_0) &= N(\mu_0, \Sigma_0) \end{aligned}$$

and find the system matrices F, G, the variance components V, W, and the prior mean μ_0 and covariance Σ_0 of the state vector at $t = 0$. (1 p)

(b) Compute the filtering distribution $p(x_t|y_{1:t})$ and the smoothing distribution $p(x_t|y_{1:T})$, for $t = 1, ..., T = 365$. For the local level parameter (α_t), plot the expected value of its filtering distribution, i.e. $\mathbb{E}(\alpha_t|y_{1:t})$. (1 p)

(c) Find the distribution of the maximum of the local slope component (β_t) over the observed period given all data. Use e.g. a histogram or kernel density estimator to present your result. [Hint: The distribution of interest is

$$p\left(\max\left\{\beta_{1:T}\right\}|y_{1:T}\right),$$

which is analytically intractable but can be computed by simulation.] (2 p)

(d) Forecast (use `dlmForecast`) the daily radiation level for the period December 5, 2013 - December 24, 2013. What is the probability that the radiation level is lower than 7 micro sieverts on Christmas Eve 2013 (December 24)? (1 p)

Good luck!

Jose and Mattias