Home

Quizzes

Theory Problems

Assignments

Assignment Questions

Video Lectures

Discussion Forums

Course Wiki

Lecture Slides

Course Schedule

Course Logistics

Course Information

Course Staff

Octave Installation

Feedback — Bayesian Priors for BNs

You achieved a score of 4.00 out of 5.00

Question 1

BDe Priors. The following is a common approach for defining a parameter prior for a Bayesian network, and is referred to as the BDe prior. Let P_0 be some distribution over possible assignments $x_1, \ldots, x_n$, and select some fixed α. For a node X with parents $\mathbf{U}$ we define $\alpha_{x|\mathbf{u}} = \alpha P_0(x, \mathbf{u})$.

For this question, assume X takes one of m values and that X has k parents, each of which takes d values. If we choose P_0 to be the uniform distribution, then what is the value of $\alpha_{x|\mathbf{u}}$?

Your Answer	Score	Explanation
$\alpha/(md^k)$ ✔	1.00	For any joint distribution it must hold that $\sum_x \sum_{\mathbf{u}} P_0(x, \mathbf{u}) = \mathbf{1}$; and for a uniform distribution all md^k terms in the sum are constant and hence must equal to $\frac{1}{md^k}$.
Total	1.00 / 1.00	

Question 2

Multiplexer CPDs. What is the form of the independence that is implied by the multiplexer CPD and that we used in our derivation of the posterior over the parameters of the simple Bayesian Network $x \to y$? (i.e. the factorization of $P(\theta_{\mathbf{X}}, \theta_{\mathbf{Y}|\mathbf{x}^1}, \theta_{\mathbf{Y}|\mathbf{x}^0} \mid \mathbf{D})$). Recall that a CPD is defined as a multiplexer if it has the strucutre $P(Y|A, Z_1, \ldots, Z_k) = \mathbf{I}\{Y = Z_a\}$ where the values of A are the natural numbers 1 through k. **Also note that the answer is specific to multiplexer CPDs and is not implied by the graph structure alone.**

Your Answer		Score	Explanation
$\theta_{\mathbf{Y}\|\mathbf{x}^0} \perp \theta_{\mathbf{Y}\|\mathbf{x}^1} \mid \mathbf{X}, \mathbf{Y}$	✔	1.00	This solution is implied by the multiplexer CPD but not by the graph structure as it relies on the CPD being independent based on which x value is active (for those with the Probabilistic Graphical Models textbook, this is similar to example 5.19 on page 174). The other options may hold, but they are implied by the graph structure and not by the specific CPD.
Total		1.00 / 1.00	

Question 3

Learning with a Dirichlet Prior. Suppose we are interested in estimating the distribution over the English letters. We assume an alphabet that consists of 26 letters and the space symbol, and we ignore all other punctuation and the upper/lower case distinction. We model the distribution over the 27 symbols as a multinomial parametrized by $\theta = (\theta_1, \ldots, \theta_{27})$ where $\sum_i \theta_i = 1$ and all $\theta_i \geq 0$. Now we go to Stanford's Green library and repeat the following experiment: randomly pick up a book, open a page, pick a spot on the page, and write down the nearest symbol that is in our alphabet. We use $X[m]$ to denote the letter we obtain in the mth experiment. In the end, we have collected a dataset $D = \{x[1], \ldots, x[2000]\}$ consisting of 2000 symbols, among which "e" appears 260 times. We use a

Dirichlet prior over θ, i.e. $P(\theta) = Dirichlet(\alpha_1, \ldots, \alpha_{27})$ where each $\alpha_i = 10$. What is the predictive probability that letter "e" occurs with this prior? (i.e., what is $P(X[2001] = "e" | D)$?) Write your answer as a decimal rounded to the nearest **ten thousandth** (0.xxxx).

0.1189

Your Answer		Score	Explanation
0.1189	✔	1.00	
Total		1.00 / 1.00	

Question 4

Learning with a Dirichlet Prior. In the setting of the previous question, suppose we had collected $M = 2000$ symbols, and the number of times "a" appeared was 100, while the number of times "p" appeared was 87. Now suppose we draw 2 more samples, X[2001] and X[2002]. If we use $\alpha_i = 10$ for all i, what is the probability of $P(X[2001] = "p", X[2002] = "a" | D)$? (round your answer to the nearest **millionth**, 0.xxxxxx)

0.002069

Your Answer		Score	Explanation
0.002069	✘	0.00	(no match)
Total		0.00 / 1.00	

Question 5

Learning with a Dirichlet Prior. In the setting of questions 3 and 4, suppose we have collected M symbols, and let $\alpha = \sum_i \alpha_i$ (we no longer assume that each $\alpha_i = 10$). In which situation(s) does the Bayesian predictive probability using the Dirichlet prior (i.e., $P(X[M+1] \mid D)$) converge to the MLE estimation for any distribution over M? You may select 1 or more options (or none of them, if you think none apply).

Your Answer		Score	Explanation
$M \to 0$ and α is fixed and non-zero	✔	0.20	The Dirichlet prior is a weighted average of the prior mean and the MLE estimate. Thus, if $M \to 0$ for a fixed (non-zero) value of α then the prior will dominate the actual counts, preventing our estimate from converging to the MLE estimation.
Both α and M are fixed and non-zero for some fixed distribution over α	✔	0.20	The Dirichlet prior is a weighted average of the prior mean and the MLE estimate. In this case, for some fixed α we can find a distribution over M that is not the same as the distribution over α. Thus, our prior will keep us some constant away from the MLE distribution.
None of the above	✔	0.20	There is at least one answer here that converges to the MLE estimation. Convergence does not mean that for any given value of α, M the estimate will be exactly the MLE estimate, but only that it becomes infinitesimally close to the MLE estimate.
✔ $\alpha \to 0$ and M is fixed	✔	0.20	The Dirichlet prior is a weighted average of the prior mean and the MLE estimate. Thus, if $\alpha \to 0$ for a fixed value of M then the probability will be dominated by the actual counts as the influence of our prior vanishes and will converge to MLE estimation.
✔ $M \to \infty$ and α is fixed	✔	0.20	The Dirichlet prior is a weighted average of the prior mean and the MLE estimate. Thus, if $M \to \infty$ for a fixed value of α then the probability will be dominated by the actual counts and will converge to MLE estimation.

Total	1.00 / 1.00