CS/CNS/EE 155: Probabilistic Graphical Models
Problem Set 1

Handed out: 8 Oct 2009
Due: 21 Oct 2009

---

1 Probability Theory

Preamble In class and in [KF09], the chain rule and Bayes' rule are cited as "the most important rules" for the probabilistic analysis of complex systems. Alternatively, some authors choose to cite the

$$\textbf{chain rule (product rule):} \quad p(X,Y) = p(Y|X)p(X)$$

and the

$$\textbf{sum rule (marginalization):} \quad p(X) = \sum_{Y} p(X,Y)$$

as the most important rules. In fact, given the product rule, Bayes' rule is readily derived from the sum rule. Similarly, the sum rule is readily derived from Bayes' rule. As stated in [Bis06], on page 839:

> [a]ll of the probabilistic inference and learning manipulations discussed in this book, no matter how complex, amount to repeated application of these two equations [the chain rule and sum rule]

(Providing scope to just how fundamental these simple equations are!).

Bishop goes on to describe how one could very well "proceed to formulate and solve complicated probabilistic models purely by *algebraic* manipulations"—but that graphical representations (the focus of this course) greatly facilitate visualization, insight into the properties, learning and inference of these complex probabilistic models.

The bottom line: the chain rule, the sum rule, and Bayes' rule underlie (nearly) all of the probabilistic analysis that we will encounter in this course. Furthermore, graphical models are tools which catalyze the application of these rules to complex systems. [1]

[1] The following problems assume some familiarity with probability theory. For additional introductory problems, read § 2.1 of [KF09], § 1.2 of [Bis06], and see the course webpage for CS228 http://www.stanford.edu/class/cs228/materials.html at Stanford, in particular http://www.stanford.edu/class/cs228/Handouts/ps0-sol.pdf.

Problems [22 points]

1. **[5 points]** Find a joint probability distribution $P(X_1, .., X_n)$ such that X_i, X_j are independent for all $i \neq j$, but $X_1, \ldots, X_n$ are not jointly independent.

2. **Conditional independence properties**

   For sets of random variables $\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{Z}$:

   (a) **[6 points]** Prove the weak union property of conditional independence:

   $$(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} | \mathbf{Z}) \implies (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}, \mathbf{W})$$

   (b) **[6 points]** Prove the contraction property of conditional independence:

   $$(\mathbf{X} \perp \mathbf{W} | \mathbf{Z}, \mathbf{Y}) \,\&\, (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) \implies (\mathbf{X} \perp \mathbf{Y}, \mathbf{W} | \mathbf{Z})$$

3. **[5 points] Valid probability distributions**

   Show that conditional parametrization based on cyclic graphs can lead to improper probability distributions.

2 Bayesian Networks

1. **Naive Bayes Model [22 points]**

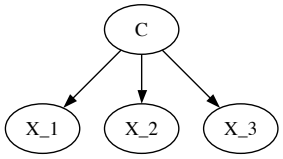

Figure 1: Bayes' net for problem **Bayesian Networks** 1

   (a) **[6 points]** Consider the naive Bayes' model given in figure 1, where C with $Val(C) = \{T, F\}$ is a binary class variable, and $\{X_1, X_2, ..., X_n\}$ are binary observations about the value of C. Derive an efficient formula for the prediction

   $$\hat{C} = \arg\max_c p(C = c | X_1, X_2, \ldots, X_n),$$

   (i.e., use the chain rule and Bayes' rule to find a more compact expression for the posterior), where the prediction is the maximum value of the posterior $p(C|X_1, X_2, \ldots, X_n)$.

   (b) **[6 points]** Suppose

   $$p(X_i = T | C = T) = 0.6 \text{ for all } i$$
   and
   $$p(X_i = T | C = F) = 0.4 \text{ for all } i$$

   Calculate, as a function of n, the ratio of posteriors

   $$p(C = T | X_1 = T = \cdots = X_n) / p(C = F | X_1 = T = \cdots = X_n)$$

(c) **[5 points]** Now suppose that we know that $X_1, X_2, \ldots, X_n$ are all identical copies of one another. What is the ratio of posteriors in this case if we take this dependency into account?

(d) **[5 points]** Based on these calculations, what conclusions can be drawn about the relationship between the prediction and the conditional independence assumption of the naive Bayes' model?

2. **D-separation [12 points]**

Preamble: d-separation Given an expression for a joint distribution in terms of a product of conditional distributions, we could test whether any potential conditional independence property holds by repeated application of the sum and chain rules. However, this approach would likely be very time consuming.

Fortunately, conditional independence properties of the joint distribution can be read directly from the graph (without any algebraic manipulations), using the framework of *d-separation*, where the "d" stands for "directed."

(a) **[7 points]** Given the Bayes' net in figure 2, which of the following conditional independence statements hold:

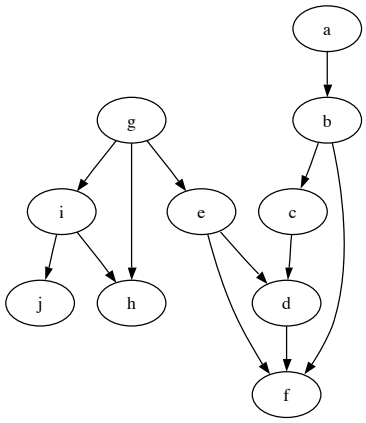

Figure 2: Bayes' net for problem **Bayesian Networks** 2a

i. $a \perp f$
ii. $a \perp g$
iii. $b \perp i \mid f$
iv. $d \perp j \mid g, h$
v. $i \perp b \mid h$
vi. $j \perp d$
vii. $i \perp c \mid h, f$

(b) **[5 points]** Assume the following conditional independencies on the random variables ordered a, b, c, d: $a \perp b$, $\neg(d \perp a)$ and $a \perp d \mid b, c$. Construct a Perfect map (Bayes' net) that satisfies these (and only these) independence assumptions.

3. **Towards inference in Bayesian Networks [22 points]**

   (a) **[7 points]** Suppose you have a Bayes' net over variables $(X_1, ..., X_n)$ and all variables except X_i are observed. Using the chain rule and Bayes' rule, find an efficient algorithm to compute $P(X_i|\{X_1, ..., X_n\} - X_i)$. In particular, your algorithm should not require evaluation of the joint distribution.

   (b) **[7 points]** Find an efficient algorithm to sample from a Bayesian network. Hint: Show that for any joint distribution $P(X, Y)$ you can sample by first drawing a sample $x \sim P(X)$ and then drawing a sample $y \sim P(Y|X = x)$.

   (c) **[8 points]** Find an efficient (randomized) procedure that, when applied to a Bayesian Network containing variable X and state $x \in Val(X)$, and for fixed parameters ϵ and δ, in only polynomially (in $1/\delta$ and $1/\epsilon$) many iterations, outputs a value p' such that with probability at least $1 - \delta$, $|P(X = x) - p'| < \epsilon$.
   Hint: take a look at Fact 1 in the "Notes for Lecture 2, CS 101.2", found at http://www.cs.caltech.edu/courses/cs101.2/slides/cs101.2-02-Bandits-notes.pdf

4. **Marginal networks [22 points]**

   (a) **[7 points]** Consider the alarm network in figure 3:

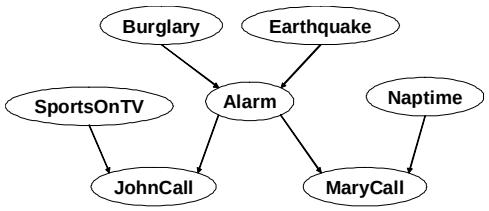

Figure 3: Bayes' Network for Problem **Bayes' Network** 4a

   Construct a Bayesian network structure with nodes Burglary, Earthquake, JohnCall, MaryCall, SportsOnTV, and Naptime, which is a minimal I-map for the marginal distribution over those variables defined by the above network. Be sure to get *all* dependencies that remain from the original network.

   (b) **[10 points]** Generalize the procedure you used above to an arbitrary network. More precisely, assume we are given a network BN, an ordering $X_1, \ldots, X_n$ that is consistent with the ordering of the variables in BN, and a node X_i to be removed. Specify a network BN' such that BN' is consistent with this ordering, and such that BN' is a minimal I-map of $P_{BN}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Your answer must be an explicit specification of the set of parents for each variable in BN'.

   (c) **[5 points]** What happens if this algorithm is used to remove the class variable of a Naive Bayes' model?

References

[Bis06] C. Bishop. *Pattern Recognition and Machine Learning.* Springer Science+Business Media, LLC, New York, NY, 2006.

[KF09] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* The MIT Press, Cambridge, MA, 2009.