bayesian-learning-exercises

March 26, 2019

1 20. Learning Probabilistic Models

20.1 [bayes-candy-exercise] The data used for Figure Section **??** on page Section **??** can be viewed as being generated by h_5. For each of the other four hypotheses, generate a data set of length 100 and plot the corresponding graphs for $P(h_i \mid d_1, \ldots, d_N)$ and $P(D_{N+1} = lime \mid d_1, \ldots, d_N)$. Comment on your results.

20.2 Repeat Exercise Section **??**, this time plotting the values of $P(D_{N+1} = lime \mid h_{\mathrm{MAP}})$ and $P(D_{N+1} = lime \mid h_{\mathrm{ML}})$.

20.3 [candy-trade-exercise] Suppose that Ann's utilities for cherry and lime candies are c_A and ℓ_A, whereas Bob's utilities are c_B and ℓ_B. (But once Ann has unwrapped a piece of candy, Bob won't buy it.) Presumably, if Bob likes lime candies much more than Ann, it would be wise for Ann to sell her bag of candies once she is sufficiently sure of its lime content. On the other hand, if Ann unwraps too many candies in the process, the bag will be worth less. Discuss the problem of determining the optimal point at which to sell the bag. Determine the expected utility of the optimal procedure, given the prior distribution from Section Section **??**.

20.4 Two statisticians go to the doctor and are both given the same prognosis: A 40% chance that the problem is the deadly disease A, and a 60% chance of the fatal disease B. Fortunately, there are anti-A and anti-B drugs that are inexpensive, 100% effective, and free of side-effects. The statisticians have the choice of taking one drug, both, or neither. What will the first statistician (an avid Bayesian) do? How about the second statistician, who always uses the maximum likelihood hypothesis?

The doctor does some research and discovers that disease B actually comes in two versions, dextro-B and levo-B, which are equally likely and equally treatable by the anti-B drug. Now that there are three hypotheses, what will the two statisticians do?

20.5 [BNB-exercise] Explain how to apply the boosting method of Chapter Section **??** to naive Bayes learning. Test the performance of the resulting algorithm on the restaurant learning problem.

20.6 [linear-regression-exercise] Consider N data points (x_j, y_j), where the y_js are generated from the x_js according to the linear Gaussian model in Equation (Section **??**). Find the values of θ_1, θ_2, and σ that maximize the conditional log likelihood of the data.

20.7 [noisy-OR-ML-exercise] Consider the noisy-OR model for fever described in Section Section **??**. Explain how to apply maximum-likelihood learning to fit the parameters of such a model to a set of complete data. (*Hint*: use the chain rule for partial derivatives.)

20.8 [beta-integration-exercise] This exercise investigates properties of the Beta distribution defined in Equation (Section **??**).

1. By integrating over the range $[0, 1]$, show that the normalization constant for the distribution beta$[a, b]$ is given by $\alpha = \Gamma(a + b)/\Gamma(a)\Gamma(b)$ where $\Gamma(x)$ is the **Gamma function**, defined by $\Gamma(x + 1) = x \cdot \Gamma(x)$ and $\Gamma(1) = 1$. (For integer x, $\Gamma(x + 1) = x!$.)

2. Show that the mean is $a/(a + b)$.

3. Find the mode(s) (the most likely value(s) of θ).

4. Describe the distribution beta$[\epsilon, \epsilon]$ for very small ϵ. What happens as such a distribution is updated?

20.9 [ML-parents-exercise] Consider an arbitrary Bayesian network, a complete data set for that network, and the likelihood for the data set according to the network. Give a simple proof that the likelihood of the data cannot decrease if we add a new link to the network and recompute the maximum-likelihood parameter values.

20.10 Consider a single Boolean random variable Y (the "classification"). Let the prior probability $P(Y = true)$ be π. Let's try to find π, given a training set $D = (y_1, \ldots, y_N)$ with N independent samples of Y. Furthermore, suppose p of the N are positive and n of the N are negative.

1. Write down an expression for the likelihood of D (i.e., the probability of seeing this particular sequence of examples, given a fixed value of π) in terms of π, p, and n.

2. By differentiating the log likelihood L, find the value of π that maximizes the likelihood.

3. Now suppose we add in k Boolean random variables $X_1, X_2, \ldots, X_k$ (the "attributes") that describe each sample, and suppose we assume that the attributes are conditionally independent of each other given the goal Y. Draw the Bayes net corresponding to this assumption.

4. Write down the likelihood for the data including the attributes, using the following additional notation:

   - α_i is $P(X_i = true | Y = true)$.
   - β_i is $P(X_i = true | Y = false)$.
   - p_i^+ is the count of samples for which $X_i = true$ and $Y = true$.
   - n_i^+ is the count of samples for which $X_i = false$ and $Y = true$.
   - p_i^- is the count of samples for which $X_i = true$ and $Y = false$.
   - n_i^- is the count of samples for which $X_i = false$ and $Y = false$.

   [*Hint*: consider first the probability of seeing a single example with specified values for $X_1, X_2, \ldots, X_k$ and Y.]

5. By differentiating the log likelihood L, find the values of α_i and β_i (in terms of the various counts) that maximize the likelihood and say in words what these values represent.

6. Let $k = 2$, and consider a data set with 4 all four possible examples of thexor function. Compute the maximum likelihood estimates of π, α_1, α_2, β_1, and β_2.

7. Given these estimates of π, α_1, α_2, β_1, and β_2, what are the posterior probabilities $P(Y = true | x_1, x_2)$ for each example?

20.11 Consider the application of EM to learn the parameters for the network in Figure Section **??**(a), given the true parameters in Equation (Section **??**).

1. Explain why the EM algorithm would not work if there were just two attributes in the model rather than three.
2. Show the calculations for the first iteration of EM starting from Equation (Section **??**).
3. What happens if we start with all the parameters set to the same value p? (*Hint*: you may find it helpful to investigate this empirically before deriving the general result.)
4. Write out an expression for the log likelihood of the tabulated candy data on page Section **??** in terms of the parameters, calculate the partial derivatives with respect to each parameter, and investigate the nature of the fixed point reached in part (c).