5 October 2010

1. (a) You obtain infinitely many copies of the text *Alice in Wonderland* and decide to play a word game with it. You cut apart each page of each copy into individual letters, throw all the letters in a bag, shake the bag, and draw three letters at random from the bag. What is the probability that you will be able to spell *tea*?

   **Answer:** From Section 2.5.2 recall that for *Alice in Wonderland*, the frequencies of the letters $e$, $t$, and $a$ are $\pi_e = 0.126$, $\pi_t = 0.099$, and $\pi_a = 0.082$, respectively. You’d need to draw one of each. There are six possible orders in which these letters could be drawn, and each order has probability $\pi_t \pi_e \pi_a = 0.00102$. Thus the total probability of being able to spell *tea* is 0.00614.

   (b) Why did the problem specify that you obtained infinitely many copies of the text? Suppose that you obtained only one copy of the text? Would you have enough information to compute the probability of being able to spell *tea*? Why?

   **Answer:** If you had only finitely many copies of the text, then the draws would no longer be conditionally independent. If your first draw were the letter $t$, for example, your chance of drawing $e$ or $a$ would increase. You’d need to know the total number of letters in all the copies of the text in order compute the exact probability of being able to spell *tea*.

2. For the constituent-order example given in Section 2.8, let $\gamma_1 = 0.6$, $\gamma_2 = 0.4$, and $\gamma_3 = 0.3$. Compute the probabilities of all six possible word orders.

   **Answer:** Consulting the table in Section 2.8, we can compute the normalizing function as

   $Z = 1 - \gamma_1 (1 - \gamma_2) \gamma_3 - (1 - \gamma_1) \gamma_2 (1 - \gamma_3)$

   $= 0.78$

   Plugging everything else in, we find that the probabilities of the six word orders are thus
3. For adult female native speakers of American English, the distribution of first-formant frequencies for the vowel [e] is reasonably well modeled as a normal distribution with mean 608Hz and standard deviation 77.5Hz. What is the probability that the first-formant frequency of an utterance of [e] for a randomly selected adult female native speaker of American English will be between 555Hz and 697Hz? (Hint: in R, the \texttt{dnorm()} function gives you access the probability density function for the normal distribution, and the \texttt{pnorm()} function gives you access to the cumulative distribution function for the normal distribution.)

\textbf{Answer:} In R:

\begin{verbatim}
> pnorm(697, 608, 77.5) - pnorm(555, 608, 77.5)
\end{verbatim}

\begin{verbatim}
[1] 0.6275673
\end{verbatim}

The probability is 0.628.

4. Use leave-one-out cross-validation to calculate the cross-validated likelihood of kernel density estimates (using a normal kernel) of male adult speaker [a] and [i] F2 formants from the Peterson and Barney dataset. Plot the cross-validated likelihood as a function of kernel bandwidth. Are the bandwidths that work best for [a] and [i] similar to each other? Show the code you wrote, and explain your results. (The dataset is available on the class website.)

\textbf{Answer:}

Below is R code to solve this problem.

\begin{verbatim}
> aa <- subset(pb.dat,Vowel=="aa" & Type=="m")
> iy <- subset(pb.dat,Vowel=="iy" & Type=="m")
> ## this is a general function, but in our case length(y)==1 always
> f <- function(x,y,bw) {
+ from <- min(y)
+ to <- max(y)
+ ## set up the grid of positions at which the kernel density is estimated
+ d <- density(x,bw=bw,n=to-from+1,from=from,to=to)
+ ## figure out which position on the grid on which the density is estimated
+ ## each observation corresponds to
+ idx <- sapply(y,function(z) which.max(z==d$x))
\end{verbatim}
(a) Cross-validated log-likelihood (b) Cross-validated log-likelihood (c) Optimal kernel density estimates for $\alpha$ for $i$

Figure 1: Solution to Exercise 4

```r
+   ## return the log-likelihood
+   sum(log(d$y[idx]))
+ }
> g <- function(x,bw) {
+   ll <- numeric(length(x))
+   for(i in 1:length(x)) {
+     training <- x[-i]
+     testing  <- x[i]
+     ll[i] <- f(training,testing,bw=bw)
+   }
+   sum(ll)
+ }
> b <- seq(30,90,by=2)
> cvll.aa <- sapply(b, function(bw) g(aa$F2,bw))
> plot(b,cvll.aa,type="l",xlab="Bandwidth",
+     ylab="Cross-validated log-likelihood for [a]",lwd=2)

> cvll.iy <- sapply(b, function(bw) g(iy$F2,bw))
> plot(b,cvll.iy,type="l",xlab="Bandwidth",
+     ylab="Cross-validated log-likelihood for [i]",lwd=2)

> b(which.max(cvll.aa))
[1] 58

> b(which.max(cvll.iy))
[1] 58
The optimal bandwidths for [a] and [i] are the same (though they are totally different from what we found for F0 frequency earlier on). This is encouraging because it suggests that similar density estimation procedures can be used for F2 frequencies regardless of the specific vowel being analyzed.

References