CS4786/5786: Machine Learning for Data Science, Spring 2015
4/7/2015: Summary of our three generative stories

In all cases, we assume that there are n customers. When we need an index variable ranging from 1 to n, we'll use the variable t.

We will further assume, for simplicity, that there is a fixed number $m = 5$ such that each customer makes exactly m purchases. When we need an index variable ranging from 1 to m, we'll use the variable q.

We will also assume that there are three products, pepsi, coke, and sprite,[1] which we'll refer to by Roman-numeral indices I=pepsi, II=coke, and III=sprite. Hence, d, the number of products, will always be 3. [Insert a joke about having D set to equal 500 instead.] When we need an index variable ranging from 1 to d, we'll use the variable ℓ.

When we need to index preference types, we'll use lower-case roman numerals, for example, the second preference type will be referred to as ϕ_{ii}. When we need an index variable ranging from 1 to the number of preference types K, we'll use the variable j.

So, when you see an upper-case Roman numeral, think "an actual product", like "coke". When you see a lower-case Roman numeral, think "a preference type", like "likely to buy either pepsi or sprite". Given these visual mnenomics, on no account will we use "i" as a variable!

Mixture of multinomials

- *Generative story*: Mother Nature has predetermined K preference types. For example, we might have $K = 4$ and ϕ_i, ϕ_{ii}, ϕ_{iii}, and ϕ_{iv} as follows:

$\phi_i[I] = .9$	$\phi_i[II] = .09$	$\phi_i[III] = .01$	(mostly likes to buy pepsi)
$\phi_{ii}[I] = .2$	$\phi_{ii}[II] = .6$	$\phi_{ii}[III] = .2$	(mostly likes to buy coke)
$\phi_{iii}[I] = .1$	$\phi_{iii}[II] = .2$	$\phi_{iii}[III] = .7$	(mostly likes to buy sprite)
$\phi_{iv}[I] = .33$	$\phi_{iv}[II] = .33$	$\phi_{iv}[III] = .34$	(no real preference among the three)

  Mother Nature has also predetermined that there is a distribution π over the K preference types. For example, it might be $\pi[i] = .6, \pi[ii] = .3, \pi[iii] = .09, \pi[iv] = .01$, that is, the most likely preference type is that of a pepsi-lover.

  To create the t^{th} customer, Mother Nature first picks their *customer type* c_t according to π. So in our example, c_t will be one of i, ii, iii, iv. Then, the t^{th} customer makes all of their m purchases according to the chosen preference type ϕ_{c_t}.

  For example, customer 1 might be allocated to the most likely type, i (probability .6 according to π); and then for their five purchases they buy five pepsis, the most likely product according to ϕ_i, and none of the other products. And the same thing might happen with the next 99 customers. Whereas customer 100 might get an unlikely assignment $c_{100} = iii$ (probability .09 according to π), and then, choosing products according to ϕ_{iii}, buy 1 coke and 4 sprites.

- *The data we are actually given:* For the t^{th} customer, we have the d-dimensional vector $\mathbf{x}_t$, where $\mathbf{x}_t[I]$ is the number of Is that that customer bought, $\mathbf{x}_t[II]$ is the number of IIs that that customer bought, and $\mathbf{x}_t[III]$ is the number of IIIs that that customer bought.

- *In real life: the task is, given the data, to recover the hidden values.*

[1] Our use of lowercase indicates that these are completely made-up product names, not having anything to do with real-life brands such as Pepsi...

Anonymous second model - changing preferences for each purchase

- *Generative story*: As in the previous story, Mother Nature has predetermined K preference types and the distribution π over the K preference types. See the examples above.

  To create the t^{th} customer's q^{th} purchase[2], Mother Nature first picks the preference $c_t[q]$ being utilized for this purchase according to π. So in our example, $c_t[q]$ will be one of i, ii, iii, iv. Then, the t^{th} customer makes their q^{th} purchase according to the chosen preference type $\phi_{c_t[q]}$.

  For example, customer first 1st purchase gets chosen by Mother Nature according to π to be motivated by the most likely preference type, i (probability .6 according to π), and by choosing according to $\phi_{c_1[1]} = \phi_i$, the customer buys a pepsi. Mother Nature also decides by a draw according to π that this customer's 2nd purchase is also motivated by preference type i, and again, choosing according to $\phi_{c_1[2]} = \phi_i$, the customer again buys a pepsi. For the third purchase, Mother Nature rolls the die and comes up with $c_1[3] = ii$, the second most likely preference type according to π. Given this preference type, the customer then decides to buy a coke (most likely purchase according to ϕ_{ii}). Customer 1's 4th and 5th purchases are made similarly.

  For customer 2, as with customer 1, the most likely outcome Mother Nature will choose as preference type for their first purchase is $c_2[1] = i$, and similarly for customer 2's other four purchases.

- *The data we are actually given:* For the t^{th} customer, we have the m-dimensional vector $\mathbf{x}_t$, where now $\mathbf{x}_t[1]$ is the product that the customer bought for their first purchase, $\mathbf{x}_t[2]$ is the product the customer bought for their second purchase, $\mathbf{x}_t[3]$ is the product that the customer bought for their third purchase, and so on.

- *In real life: the task is, given the data, to recover the hidden values.*

Latent Dirichlet allocation (LDA) . Data and task are the same as above.

- *Generative story*: Mother Nature predetermines a Dirichlet prior on possible "profiles" π — multinomials over preference types, as discussed in class. In other words, she predetermines the parameter vector $(\alpha_1, \alpha_2, \ldots, \alpha_K)$ for the Dirichlet profiles prior.
  Then, seeing no reason not to re-use a technique, she predetermines a Dirichlet prior on the preference types ϕ themselves, since they are also multinomials (but over products). This means that she predetermines the parameter vector $(\beta_1, \beta_2, \ldots, \beta_d)$ for the preference-type prior. She then predetermines the K preference types ϕ_j by drawing K samples from the preference types prior.

  To create the t^{th} customer, Mother Nature first picks this particular user's π_t, i.e., their distribution over preferences, according to the profiles prior. Then, to motivate the t^{th} customer's q^{th} purchase, Mother Nature first picks the preference $c_t[q]$ being utilized for this purchase according to π_t. So in our example, $c_t[q]$ will be one of i, ii, iii, iv. Then, the t^{th} customer makes their q^{th} purchase according to the chosen preference type $\phi_{c_t[q]}$.

  Thus, we might have π_1 be like our π in the example above — heavily weighted towards always being a pepsi lover; whereas for customer 2, it might turn out that they get $\pi_2[i] = .25, \pi_2[ii] = .25, \pi_2[iii] = .25, \pi_2[iv] = .25$, meaning that they are equally likely to be motivated by any one of the four preference types for each individual purchase.

[2]We use red to highlight changes from the preceding model.