Alin Tomescu, http://people.csail.mit.edu/~alinush
6.867 Machine learning | Prof. Tommi Jaakkola | Week 11, Tuesday, November 12th, 2013| Lecture 19

Lecture 19

Final will cover all material with emphasis on the 2nd part. If you did very bad on the midterm and you do better on the section of the final with midterm topics, then they will discount the midterm.

Independence graphs: They specify independence properties on the variables.

Why are we focusing on independence instead of dependence?

- Independence is a qualitative statement. Dependence is a quantitative statement: you can have strong / weak dependence.
- The main reason is if you know some things are independent you will have an easier time computing your model.
- Once you claim that things are independent you are imposing strong constraints on the model.

Example:

$x_i \in \{0,1\}, P(x_1, \ldots, x_n)$ we need $2^n - 1$ parameters to specify this distribution. **Why?**

If they are all independent: $P(x_1) \ldots P(x_2)$, I need n parameters.

Independence

Marginal independence

Two random variables X and Y, are marginally independent means:

$$X \perp Y \Leftrightarrow P(x, y) = P(x)P(y)$$

- LHS is an independence statement
- RHS is a factorization

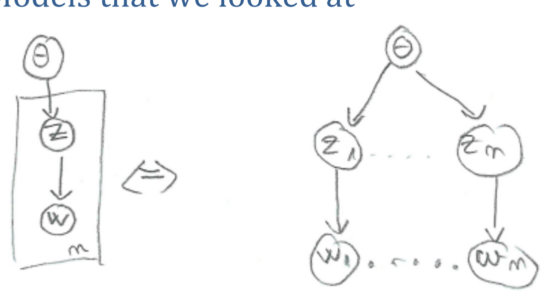
Example: First flip and second flip of a coin

Conditional independence

X and Y are conditionally independent given Z: $X \perp Y \mid Z \Leftrightarrow P(x, y \mid z) = P(x|z)P(y|z), \forall z, P(z) > 0$

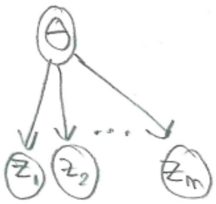
Example: Three coin flips where the first two are conditioned on the third

Models that we looked at



What independence properties does this model satisfy?

- All the z_i's are conditionally independent given θ (once it is given a value)
- All the w_i's are conditionally independent given θ
  - Once I know θ, each word is sampled independently
- w_i's are conditionally independent given $z_1, \ldots, z_n$
- $w_i \perp \theta \mid z_i, \forall i = 1, \ldots, n$



$z_1, z_2, \ldots, z_n$ are independent given θ.

$$P(z_1, z_2, \ldots, z_n, \theta) = P(z_1, z_2, \ldots, z_n | \theta) P(\theta) \triangleq \left(\prod_{i=1}^{n} P_i(z_i|\theta) \right) P(\theta)$$

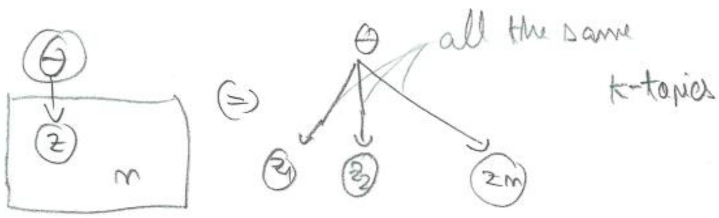
- The P_i notation is used in order to make it clear that $P_i(z_i|\theta)$ does not have to be equal to $P_j(z_j|\theta)$

We also have **exchangeability** here: $z_1, \ldots, z_n$ are exchangeable, since if I integrate/marginalize over theta, then:

$$P(z_1 = l_1, z_2 = l_2, \ldots, z_n = l_n) = P(z_1 = l_i, z_2 = l_j, \ldots, z_n = l_k)$$

Where we permuted the values of l_i

Key idea: Graph separation ⇒ independence



$$P(z_i = l \mid z^{-i}) = P(z_i = l \mid z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$$

$$\theta \sim Dirichlet\left(\frac{\alpha}{k}, \ldots, \frac{\alpha}{k}\right)$$

$$P(z_i = l \mid z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n) = \frac{\frac{\alpha}{k} + n^{-i}(l)}{\alpha + n - 1}$$

Let the number of topics be infinite:

$$\lim_{k\to\infty} \frac{\frac{\alpha}{k} + n^{-i}(l)}{\alpha + n - 1} = \frac{n^{-i}(l)}{\alpha + n - 1}$$

What is the probability of choosing the first?

$P(z_1 = l) = 0$ as k goes to infinity

$$P(z_i = l_{new}) \xrightarrow[k\to\infty]{} \frac{\alpha}{\alpha + n - 1}$$

Where l_{new} is a value that has not been chosen yet as a topic before choosing topic i.

Chinese restaurant process

This shows the clustering effect explicitly.
Restaurant has infinitely many tables $k = 1, \ldots$.
Customers are indexed by $i = 1, \ldots$, with values ϕ_i
Tables have values θ_k drawn from G_0

K = total number of occupied tables so far.
n = total number of customers so far.
n_k = number of customers seated at table k

Generating from a CRP:
customer 1 enters the restaurant and sits at table 1.
$\phi_1 = \theta_1$ where $\theta_1 \sim G_0$, $K = 1$, $n = 1$, $n_1 = 1$
for $n = 2, \ldots$,

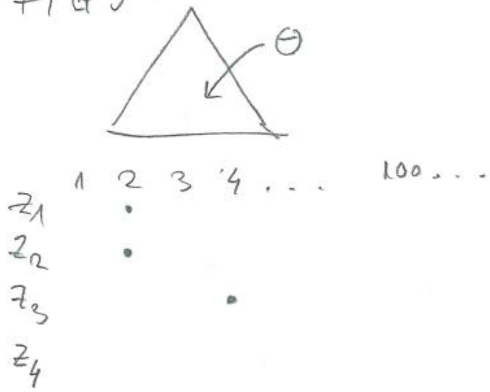
customer n sits at table $\begin{cases} k & \text{with prob } \frac{n_k}{n-1+\alpha} \quad \text{for } k = 1 \ldots K \\ K+1 & \text{with prob } \frac{\alpha}{n-1+\alpha} \quad \text{(new table)} \end{cases}$

if new table was chosen **then** $K \leftarrow K + 1$, $\theta_{K+1} \sim G_0$ **endif**
set ϕ_n to θ_k of the table k that customer n sat at; set $n_k \leftarrow n_k + 1$
endfor

Clustering effect: New students entering a school join clubs in proportion to how popular those clubs already are ($\propto n_k$). With some probability (proportional to α), a new student starts a new club.

What is this? http://en.wikipedia.org/wiki/Chinese_restaurant_process

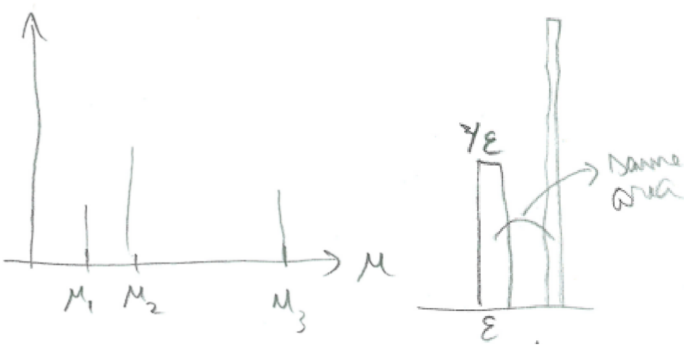
FIG 5



Gaussian mixture models

$$P(x|\theta) = \sum_{z=1}^{k} P_z N(x; \mu_z, \sigma^2)$$

Where σ is fixed for simplicity.



Spikey densities

Dirac function $\delta_{\mu_z}(\mu) = \begin{cases} 1 \text{ or } \infty \text{ (not sure)}, \mu_z = \mu \\ 0, \mu_z \neq \mu \end{cases}$

$$\tilde{\varsigma}(\mu) = \sum_{z=1}^{k} P_z \delta_{\mu_z}(\mu)$$

$$\int \delta_{\mu_z}(\mu) d\mu = 1$$

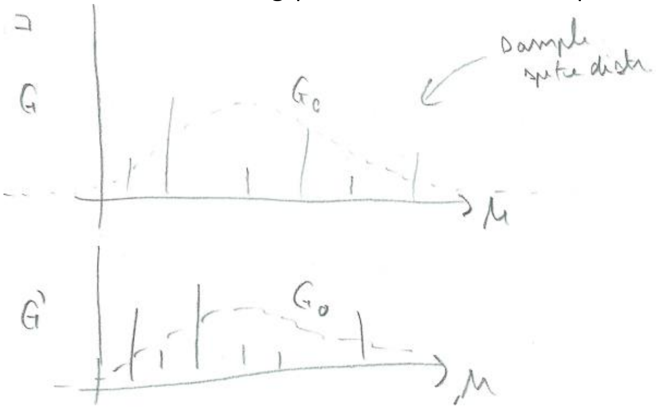
$$\int f(\mu) \delta_{\mu_z}(\mu) d\mu = f(\mu_z)$$

$$P(x|\theta) = E_{\mu \sim \tilde{\varsigma}}\{N(x; \mu, \sigma^2)\}$$

Finding that spikey density

Define a random variable called a Dirichlet process.

$$G \text{ is a pdf}, G \sim DP(\alpha, G_0)$$



k;