Alin Tomescu, http://people.csail.mit.edu/~alinush

6.867 Machine learning | Prof. Tommi Jaakkola | Week 9, Thursday, October 31st, 2013| Lecture 16

Lecture 16

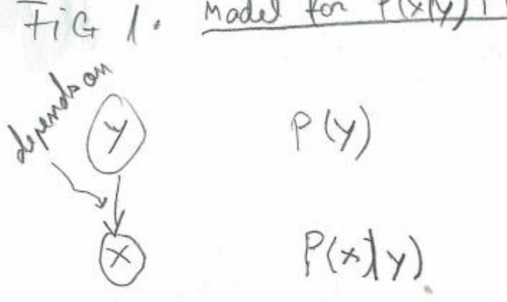
Project grading criteria: How you describe what you achieved, show what you understand (practical / theoretical aspect of the material). It does not matter if you succeeded in your classification task

2 topics that will be on final: SVM with kernels, EM algorithm.

$$P(x,y;\theta) = \sum_{y=1}^{k} P(x \mid y;\theta)\, P(y;\theta)$$

A simple graphical model for this will look like:

FIG 1. Model for P(x|y) P(y)

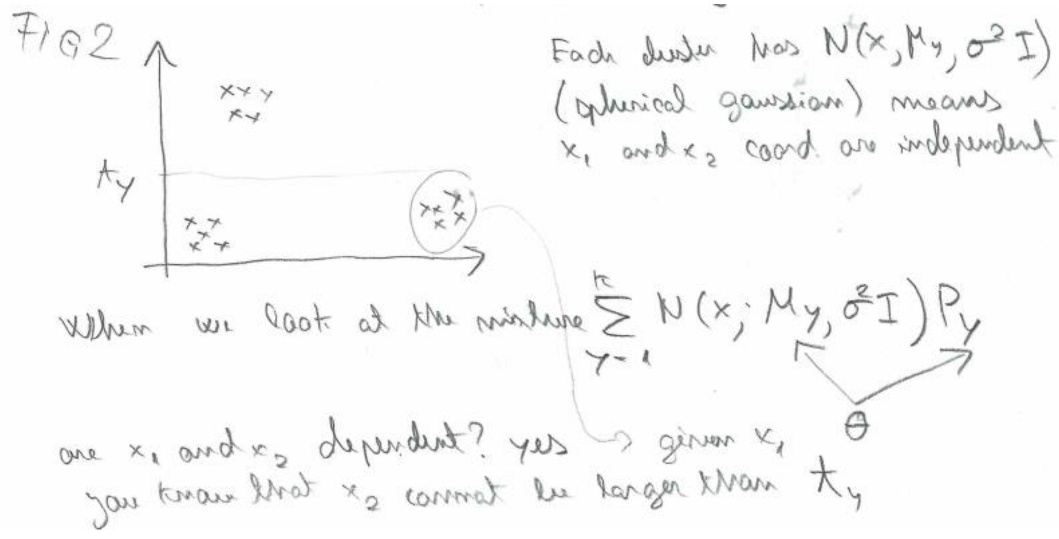


This is a "compact description for generating the values" and this is how you sample/generate values:

- y does not depend on anything, so just sample
- x depends on y, so you can only sample it after you sampled an y

The EM algorithm

FIG 2



Each cluster has $N(x, \mu_y, \sigma^2 I)$ (spherical gaussian) means x_1 and x_2 coord are independent

When we look at the mixture $\sum_{y=1}^{k} N(x; \mu_y, \sigma^2 I)\, P_y$

are x_1 and x_2 dependent? yes → given x_1 you know that x_2 cannot be larger than t_y

First step: Initialize. We will see this is quite important, since it defines a starting point from which we start refining.

$$\theta^{[0]}$$

E-step: Performs the *data completion.* In each step, we map each data point $x^{(i)}$ to k different data points $\left(x^{(i)}, y\right)$ but not with equal weighting (see previous q function).

$$q^{[m]}(y \mid i) = P\left(y \mid x^{(i)}; \theta^{[m]}\right) = \frac{P\left(x^{(i)} | y; \theta^{[m]}\right) P_y^{[m]}}{P\left(x^{(i)}; \theta^{[m]}\right)} = \frac{P\left(x^{(i)} | y; \theta^{[m]}\right) P_y^{[m]}}{\sum_{y'=1}^{k} P(x^{(i)} \mid y'; \theta^{[m]}) P_{y'}^{[m]}}$$

We do this for all $y = 1, \dots, k$ and for the entire data set $i = 1, \dots, n$.

Note: Now we kind of have nk points after the $x^{(i)} \to \left(x^{(i)}, y\right)$ mapping.

M-step: Simply takes that data set and maximizes the log-likelihood of it.

$$\theta^{[m+1]} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{y=1}^{k} q^{[m]}(y \mid i) \log\left[N\left(x^{(i)}; \mu_y^{[m+1]}, \sigma^2 I\right) P_y^{[m+1]}\right], \theta = \left[\mu_y^{[m+1]}, P_y^{[m+1]}\right]$$

If we compute the argmax, we obtain:

$$\hat{p}_y^{[m+1]} = \frac{\sum_{i=1}^{n} q^{[m]}(y|i)}{n}, i = 1 \dots k$$

$$\hat{\mu}_y^{[m+1]} = \frac{1}{\sum_{i=1}^{n} q^{[m]}(y \mid i)} \sum_{i=1}^{n} q^{[m]}(y \mid i) \cdot x^{(i)}, i = 1 \dots k$$

Note: $\sum_{y=1}^{k} q(y \mid i) = 1$

Analyzing the EM algorithm

Let's see how starting from a poor model might lead to improvement and how important the initialization step is.

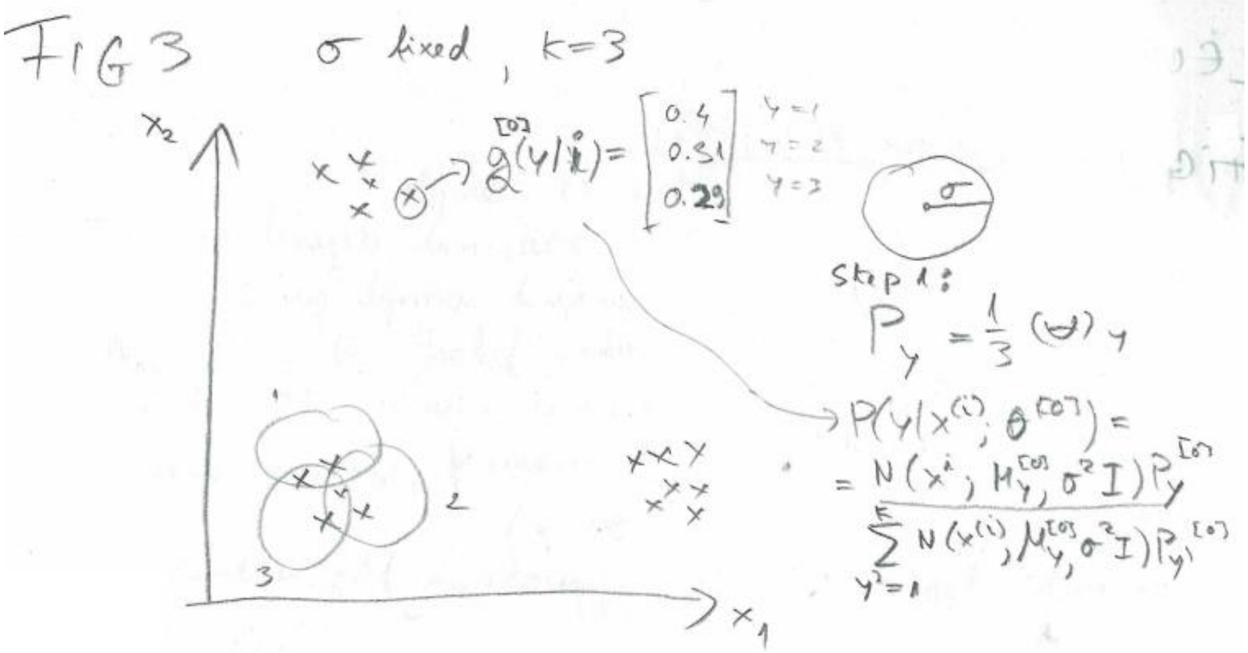


FIG 3
σ fixed, k=3
x_2
$q^{[0]}(y|i) = \begin{bmatrix} 0.4 \\ 0.31 \\ 0.29 \end{bmatrix}$
y=1
y=2
y=3
σ
Step 1:
$P_y = \frac{1}{3} (\forall) y$
$P(y|x^{(i)}; \theta^{[0]}) = \frac{N(x^i, \mu_y^{[0]}, \sigma^2 I) P_y^{[0]}}{\sum_{y'=1}^{k} N(x^{(i)}, \mu_{y'}^{[0]}, \sigma^2 I) P_{y'}^{[0]}}$
1
2
3
x_1

Initialization step: Pick probabilities for each y, pick mean for the Gaussians for each y.

$$P_y = \frac{1}{3}, \forall y$$

$$\mu_y = \cdots$$

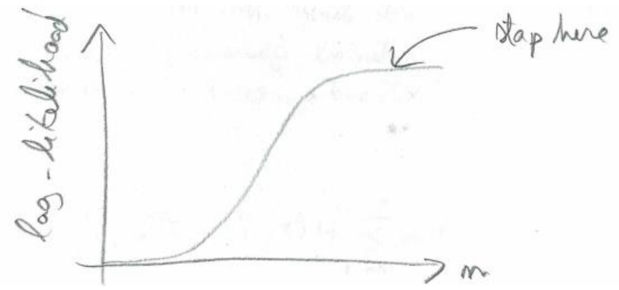
To pick the means we could just pick random points for each cluster.

E-step: compute $q^{[m]}(y \mid i) = P\left(y \mid x^{(i)}; \theta^{[m]}\right)$

$$q^{[0]}(y \mid i) = P\left(y \mid x^{(i)}; \theta^{[0]}\right)$$

M-step: In the above example, the Gaussian $y = 1$ only looks at $q(1, i)$ and goes towards the points that have higher $q(1, i)$ values. The other Gaussians proceed similarly.

Question: When do you stop? **Answer:** It generally takes an infinite amount of steps to converge to the maximum likelihood. The likelihood grows slowly initially and then it starts increasing as the Gaussians "see" the points and then it starts growing slowly again.



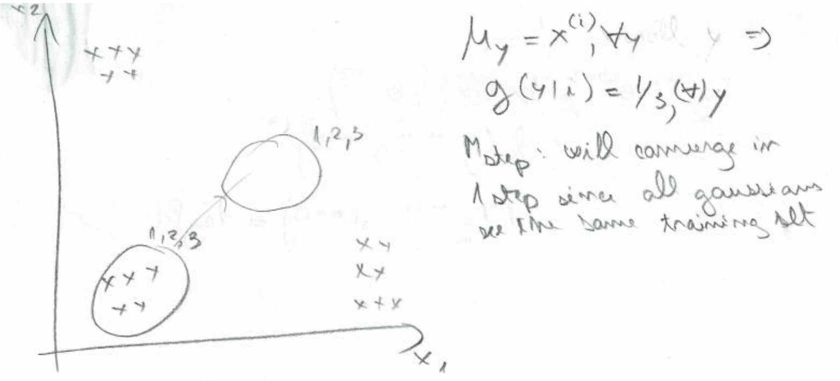
Alin Tomescu, http://people.csail.mit.edu/~alinush
6.867 Machine learning | Prof. Tommi Jaakkola | Week 9, Thursday, October 31st, 2013| Lecture 16

Note: You will always converge in the EM algorithm, but not necessarily to the best solution. Only guaranteed to converge to a locally optimal solution.

Question: What will happen with different initialization, why is it important?

Case 1: $\mu_y = x^{(i)}, \forall y$ and $P_y = p, \forall y$ and fixed σ

In this case the EM algorithm will converge in one step to a non-optimal solution.


x_2
x_1
1,2,3
1,2,3
$\mu_y = x^{(i)}, \forall y \Rightarrow$
$g(y|i) = 1/3, (\forall) y$
M step: will converge in 1 step since all gaussians see the same training set

Case 2: $\mu_y = x^{(i)}, \forall y$ and $P_{y^{(i)}} \neq P_{y^{(j)}}$ and fixed σ

In this case, the same solution is computed.

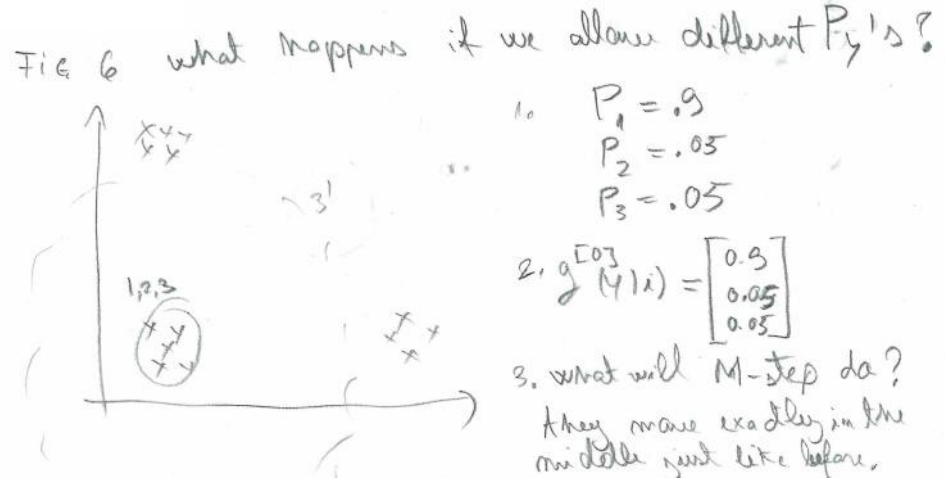


Fig 6 what happens if we allow different P_y's?
1,2,3
1. $P_1 = .9$
$P_2 = .05$
$P_3 = .05$
2. $g^{[0]}(y|i) = \begin{bmatrix} 0.9 \\ 0.05 \\ 0.05 \end{bmatrix}$
3. what will M-step do?
they move exactly in the middle just like before.

Case 3: $\mu_y = x^{(i)}, \forall y$ and $P_{y^{(i)}} \neq P_{y^{(j)}}$ and different σ

Note: These are not the only ways to initialize badly. Plenty of other bad ways…

If I allow different σ's, the Gaussian with the higher σ will "see" points further away "better" and will move towards these points more than the other Gaussians

EM algorithm as a maximization problem

We will show that the E-step and M-step both maximize for the log-likelihood.

What is this lower bound that I am constantly maximizing and how it relates to the log-likelihood?

If we can maximize $P\left(x^{(i)};\theta\right)$, we can also maximize over the sum:

$$P(x^{(i)};\theta) = \sum_{y=1}^{k} P\left(x^{(i)},y;\theta\right)$$

$$P\left(x^{(i)},y;\theta\right) = P\left(x^{(i)} \mid y;\theta\right)P(y;\theta) = P\left(y \mid x^{(i)};\theta\right)P\left(x^{(i)};\theta\right)$$

$$l(q,\theta) = \sum_{y=1}^{k} q(y \mid i) \log P\left(x^{(i)},y;\theta\right) + \sum_{y} q(y \mid i) \log \frac{1}{q(y \mid i)}$$

We added this (Shanon) entropy term (measure of uncertainty) $\sum_y q(y \mid i) \log \frac{1}{q(y \mid i)}$ as a penalty.

Claim 1: For any parameters, $l(q,\theta) \leq \log P\left(x^{(i)};\theta\right)$

Claim 2: If I start from parameters $\theta^{[m]}$, $l\left(q^{[m]},\theta^{[m]}\right) = \log P\left(x^{(i)};\theta^{[m]}\right)$, when $q^{[m]}(y \mid i) = P\left(y \mid x^{(i)};\theta^{[m]}\right)$

$$l\left(q^{[m]},\theta^{[m]}\right) = \sum_{y=1}^{k} q^{[m]}(y \mid i) \left(\frac{\log P\left(x^{(i)},y;\theta^{[m]}\right)}{q^{[m]}(y \mid i)}\right) = \sum_{y=1}^{k} q^{[m]}(y \mid i) \left(\frac{\log P\left(x^{(i)};\theta^{[m]}\right)P\left(y|x^{(i)};\theta^{[m]}\right)}{q^{[m]}(y \mid i)}\right)$$
$$= \sum_{y=1}^{k} q^{[m]}(y \mid i) \left(\frac{\log P\left(x^{(i)};\theta^{[m]}\right)q^{[m]}(y \mid i)}{q^{[m]}(y \mid i)}\right) = \sum_{y=1}^{k} q^{[m]}(y \mid i) \log P\left(x^{(i)};\theta^{[m]}\right) = \log P\left(x^{(i)};\theta^{[m]}\right)$$

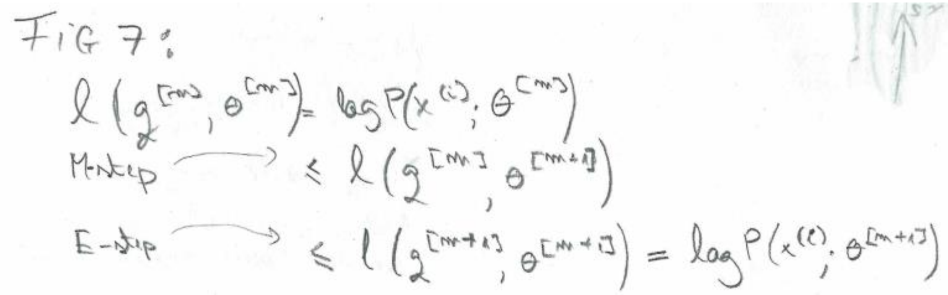


FIG 7:

$$l\left(q^{[m]},\theta^{[m]}\right) = \log P\left(x^{(i)};\theta^{[m]}\right)$$
$$\text{M-step} \longrightarrow \leq l\left(q^{[m]},\theta^{[m+1]}\right)$$
$$\text{E-step} \longrightarrow \leq l\left(q^{[m+1]},\theta^{[m+1]}\right) = \log P\left(x^{(i)},\theta^{[m+1]}\right)$$

$$l\left(q^{[m]},\theta^{[m]}\right) = \log P\left(x^{(i)};\theta^{[m]}\right) \leq l\left(q^{[m]},\theta^{[m+1]}\right) \leq l\left(q^{[m+1]},\theta^{[m+1]}\right) = \log P\left(x^{(i)};\theta^{[m+1]}\right)$$