Probabilistic graphic model

1. From data point of view:

   vb and kb: independent of data

   graphic: depend on each other

graphical: intuitive way of representing and visualizing the relationships between variables (conditional independence) vector-based, kernel-based: statistical view (coordinate transformation, variance) on distribution of data. vector-based: dimension reduction. kernel-based: projection, non-linear and coordinate transformation

2. (他真的是随缘在讲)

   For $maxP(x,y)$, we need to estimate label for each single x.

   $maxP(x_i,y_k) \rightarrow \prod^i \sum^k (\beta)$, where β is a k-th polynomials. That means for each x there may exists k labels y, and we need to consider all the x_i by taking product.

   It can get some local solutions which can be used as initial values or lower bound for optimal.

   It is the target. Only part of variables interest us. It is prerequisite for other tasks.
   $P(H|X) \propto P(X|H)P(H)$ and finding $Q(H)$ is very import in EM algorithm. Sometimes, it is used to eliminate the influence of hidden variables.

3. Assume labels are outstanding for each class:

   From a Bayesian point of view: label equals to constrain which is ideal condition or truth and it should be as universal as possible. If label is ambiguous, that means it will vary from each individual. This will make it difficult to learn.

4. (not for sure)$P(x,y) = P(y|x)P(x)$

   A knowledge-based ontology is a universal ontology which is from the understanding point of view. By using the prior structure, a distribution of data can be guaranteed.

5. Everything is a marginal estimation.

   From statistical view, latent variables are unobserved, usually indicating complex distribution. In terms of training, both latent variables and model parameters are supposed to be found, which is naturally more difficult than models with no hidden variables.

   Most of such kind of problems are solved by EM algorithm which does not guarantee a global solution. The hidden variables are obtained from expectation which is only an approximation method and not accurate enough. The problem is non-convex.

6. Key Assumption: conditional independence + Markov property

For HMM, The assumption is that given H_{t-1}, H_t is conditionally independent of H_{t-2},H_{t-3}, ... i.e. $P\left(H_t|H_{t-1},H_{t-2},\ldots,H_1\right) = P\left(H_t|H_{t-1}\right)$. This simplification turns the varied and high-cost computation into a matrix multiplication operation, which dramatically drop the computational complexity. From $O(K^T)$ to $O(TK^2)$

7. Because the model in maximizing step is not guaranteed to be convex. To illustrate, $\theta = \arg\max_\theta E_{\sim Q}\left[\log \frac{P(X,H;\theta)}{Q(H)}\right]$ and model $P(X, H; \theta)$ may not be convex and either does the expectation function E. EM is guaranteed to converge to a point with zero gradient which may not be a local minimal (saddle point) and let alone the global optimal.

   Jensen Inequality & coordinate ascend

   Jensen Inequality illustrates that $f(E(x)) \leq E(f(x))$ is valid if $f(x)$ is convex. But the model f here is not guaranteed to be convex. $f(x) = \log g(x)$ and while log is convex, $g(x)$ may not be convex. Therefore, EM algorithm based on Jensen Inequality and coordinate ascend does not guarantee a global solution.

8. K-means is an instance of EM algorithm, which converges to a zero gradient point. This property only guarantees a local optimal solution. Since K-means, basing on Gaussian Distribution assumption, is the special case of EM algorithm, it cannot cover all the cases of the given dataset. i.e. There may be some better distribution to model the problem which is the true global optimum. Even based on Gaussian Assumption, different initial data centers will arrive in different local optimal solutions.

后面这几道题建议自己先去查查前向算法后向算法，y老师讲的太玄了

9. HMM falls into a subclass of Bayesian Network named Dynamic Bayesian Network, i.e. the joint inference of a series hidden states can be written as $P\left(H_{1:t}\right) = P\left(H_1\right) P\left(H_2|H_1\right) \cdots P\left(H_t|H_{t-1}\right)$.

   Forward:try to propagate product, update the prior as well-defined as possible.

   Backward: get all data, update the order to determine a best sequence.

10. $\alpha_t(i) \equiv P\left(O_{1:t}, h_t = H_i\right)$
    $\beta_t(i) \equiv P\left(O_{t+1:T}|h_t = H_i\right)$

    and,

$$
\begin{aligned}
P\left(h_t = H_i|O_{1:T}\right) &\propto P\left(O_{1:T}|h_t = H_i\right) P\left(h_t = H_i\right) \\
&\propto P\left(O_{1:t}|h_t = H_i\right) P\left(O_{t+1:T}|h_t = H_i\right) P\left(h_t = H_i\right) \\
&\propto P\left(O_{1:t}, h_t = H_i\right) P\left(O_{t+1:T}|h_t = H_i\right)
\end{aligned}
$$

    Therefore,

    $P\left(h_t = H_i|O_{1:T}\right) = \frac{\alpha_t(i)\beta_t(i)}{\sum_i \alpha_t(i)\beta_t(i)}$

11. Let the priors be consistent and big. By applying multiplication, all the priors must be as big as possible to faster the inference(Otherwise will cancel others).

    From structural point of view, we want to use more converging products(?)

12. By calculating $\alpha \cdot \beta$, where the value changes dramatically is the import node.

13. Forward process is to update the prior which is more important for human.

14. (讲了一堆，记了个大概意思,感觉不太对。。)Human: take advantage of priors, do not need whole sequence.

    Machine: Have to exhaust all the possible results. Can only use data to maximize, to regression.

15. ?A better likelihood, 贝叶斯公式吧

16. • Most edges have zero weights. i.e. sparse adjacent matrix. (Pruning) • Applying SVD in dimension reduction. Simplify the computation.

    (后面这些什么高斯graphic model 的我放弃了。听不懂他在说什么，听懂了我觉得也跟题意没啥关系，反正10选7，合理放弃。。。)

17.
18. p
19. p
20. p