Bayesian Methods for Machine Learning

Andrey de Aguiar Salvi

December 2020

1 Class 1

Three principles:

- use prior knowledge
- choose the answer that explains the observations the most
- avoid making extra assumptions

1.1 Variable Independence

$$P(X,Y) = P(X)P(Y) \tag{1}$$

1.2 Conditional Probability

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{2}$$

where $P(X,Y)$ = joint probability and $P(Y)$ = marginal probability.

1.3 Chain Rule

$$P(X,Y) = P(X|Y)P(Y) \tag{3}$$

$$P(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z) \tag{4}$$

$$P(X_1,...,X_N) = \prod_{n=1}^{N} P(X_i|X1,...,X_{i-1}) \tag{5}$$

1.4 Marginalization

$$p(X) = \int_{-\inf}^{in} p(X,Y)dY \tag{6}$$

1.5 Bayes Theorem

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{7}$$

where $P(\theta|X)$ = posterior probability, $P(X|\theta)P(\theta)$ = Likelihood, and $P(X)$ = Evidence.

2 Class 2

2.1 Statistic Approaches

- **Frequentist**:
  - deterministic
  - θ is fixed, X is random
  - work whether data points is higher than the parameters - $|X| >> |\theta|$
  - Train models with Maximum Likelihood:

$$\hat{\theta} = arg\max_{\theta} P(X|\theta) \tag{8}$$

    to maximize the probability of data given the parameters

- **Bayesing**:
  - subjective
  - θ is random, X is fixed (given a set of forces θ, tossing a coin always gives the same result X)
  - work with data on any size - $|X|$
  - Train models with Na"(i)ve Bayes to maximize the probability of the parameters given the data

2.2 On-line learning

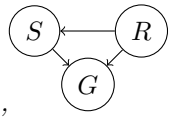
Use the current mini-batch (posterior) to update parameters, and then use it as prior in the new mini-batch.

3 Class 3

3.1 Bayesian Net

Is not a bayesian neural network. The **Nodes** are random variables and the **Edges** are the direct impact. **Model**: is the joint probability over all probabilites

$$P(X_1, ..., X_N) = \prod_{n=1}^{N} P(X_i|P_a(X_i)) \tag{9}$$

where $P_a(X_i)$ is the probability of the parent nodes from the Bayesian Net. *E.g.,* 

where R is father from G and S, S is parent from G, and the equation bellow is the respective bayesian probability

$$P(S, R, G) = P(G|S, R)P(S|R)P(R) \tag{10}$$

4 Class 5

4.1 Univariate Normal Distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

4.2 Multivariate Normal Distribution

$$\mathcal{N}(x|\mu, \Sigma^2) = \frac{1}{\sqrt{2\pi\Sigma^2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^-1(x-\mu)} \tag{12}$$

4.3 Linear Regression

$$L(w) = \sum_{i=1}^{N} (w^T x_i - y_i)^2 = ||w^T X - y|| \rightarrow \min_w \tag{13}$$

$$\hat{w} = arg\min_w L(w) \tag{14}$$

where L is the loss from the bayesian net, w are the weights and X are the data, both are parents of y (target)

$$P(w, y|X) = P(y|X, w)P(w) \tag{15}$$

$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 \mathcal{I}) \tag{16}$$

$$P(w) = \mathcal{N}(w|0, \gamma^2 \mathcal{I}) \tag{17}$$

$$P(w|y, x) = \frac{P(y, w|X)}{P(y|X)} \rightarrow \max_w \tag{18}$$

$$P(y, w|x) = \frac{P(y|x, w)}{P(w)} \rightarrow \max_w \tag{19}$$

5 Class 6 - Maximum a Posteriori

To learn a distribution

$$\theta_{MP} = arg\max_\theta P(\theta|X) = arg\max_\theta \frac{P(\theta|X)P(\theta)}{P(X)} = arg\max_\theta P(\theta|X)P(\theta) \tag{20}$$

We eliminate the denominator from the last equation, once time it don't have θ Problems:

- it is not variant to reparametrization. Ex learning about a gaussian will be not usefull in sigmoid(gaussian)
- strange "loss function"
- we do not have the posteriori of θ
- can't compute credible intervals

6 Class 7 - Conjugate Distribution

It is a way to avoid computing the evidence ($P(X)$), which is costly. In the Bayes Probability

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{21}$$

in the likelihood, $P(X|\theta)$ is fixed by the model, $P(X)$ is fixed by the data, and $P(\theta)$ is our own choice. The prior $P(\theta)$ is conjugate to the likelihood if the prior and the posterior $P(X|\theta)$ lie in the same family distributions.

*E.g.,*if the prior is $P(X|\theta) = \mathcal{N}(x|\mu, \sigma^2)$ and the posterior is $P(\theta) = \mathcal{N}(\theta|m, s^2)$, thus the conjugate of posterior $P(\theta|X)$ is $\mathcal{N}(\theta|a, b^2)$.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{\mathcal{N}(x|\theta, 1)\mathcal{N}(\theta|0, 1)}{P(X)} \tag{22}$$

$$P(\theta|X) \propto e^{-\frac{1}{2}(X-\theta)^2} e^{-\frac{1}{2}\theta^2} \tag{23}$$

$$P(\theta|X) \propto e^{-(\theta - \frac{X}{2})^2} \tag{24}$$

$$P(\theta|X) = \mathcal{N}(\theta|\frac{X}{2}, \frac{X}{2}) \tag{25}$$

7 Class 8

7.1 Gamma Distribution

$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma} \tag{26}$$

where

- $\gamma, a, b > 0$
- $\Gamma(n) = (n-1)!$
- the expectation, or mean, $\mathbb{E}[\gamma] = \frac{a}{b}$
- $Mode[\gamma] = \frac{a-1}{b}$
- $Var[\gamma] = \frac{a}{b^2}$

7.2 Precision

$$\gamma = \frac{1}{\sigma^2} \tag{27}$$

If we replace the variance in the Normal Distribution to the inverse of the Precision, we get

$$\mathcal{N}(x|\mu,\gamma^{-1}) = \frac{\sqrt{\gamma}}{\sqrt{2\pi}}e^{-\gamma\frac{(x-\mu)^2}{2}} \tag{28}$$

thus, the conjugate prior in respect to the precision is

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}}e^{-b\gamma} \tag{29}$$

$$P(\gamma) \propto \gamma^{a-1}e^{-b\gamma} \tag{30}$$

$$P(\gamma|X) = \Gamma(\gamma|a,b) \tag{31}$$

then, the prior is

$$P(\gamma|X) = \Gamma(\gamma|a,b) \propto \gamma^{a-1}e^{-b\gamma} \tag{32}$$

the posterior is

$$P(\gamma|X) \propto P(X|\gamma)P(\gamma) \tag{33}$$

dropping out all the constants, we have

$$P(\gamma|X) \propto \left(\gamma^{\frac{1}{2}}e^{-\gamma\frac{(X-\theta)^2}{2}}\right)\left(\gamma^{a-1}e^{-b\gamma}\right) \tag{34}$$

re-arranging the terms, we get

$$P(\gamma|X) \propto \gamma^{\frac{1}{2}+a-1}e^{-\gamma\frac{(X-\theta)^2}{2}}e^{-\gamma(b+\frac{(X-\mu)^2}{2})} \tag{35}$$

finally

$$P(\gamma|X) = \Gamma(a+\frac{1}{2}, b+\frac{(X-\mu)^2}{2}) \tag{36}$$

8 Class 9

8.1 Beta Distribution

$$\beta(X|a,b) = \frac{1}{\beta(a,b)}X^{a-1}(1-X)^{b-1} \tag{37}$$

where:

- $X \in [0,1]$
- $a, b > 0$

- $$\beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{38}$$
- $$\mathbb{E}[X] = \frac{a}{a+b} \tag{39}$$
- $$Mode[X] = \frac{a-1}{a+b-2} \tag{40}$$
- $$Var[X] = \frac{ab}{(a+b)^2(a+b-1)} \tag{41}$$

*E.g.,*to model a distribution with mean 0.8 and standard deviation of 0.1, thus $\mathbb{E}[X] = 0.8$ and $Var[X] = 0.1^2$. Consequently, a Beta distribution with $a = 12$ and $b = 3$ model it.

8.2 Bernoulli Distribution

The Beta distribution is the conjugate of the Bernoulli likelihood. *E.g.,*a Bernoulli likelihood from a dataset

$$P(X|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{42}$$

where N_1 is the number of ones in X and N_0 the number of zeros. Thus, the Beta distribution is

$$P(\theta) = \beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1} \tag{43}$$

Multiplying the likelihood by the prior

$$P(\theta|X) \propto P(X|\theta)P(\theta) \tag{44}$$

using the before, but replacing the terms by the equivalent equations above, we have

$$P(\theta|X) \propto \theta^{N_1}(1-\theta)^{N_0}\beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1} \tag{45}$$

and rearranging the terms, we have

$$P(\theta|X) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \tag{46}$$

and finally, recognizing the equation before as a Beta distribution, we have

$$P(\theta|X) = \beta(N_1+a, N_0+b) \tag{47}$$

8.3 Posteriors

- Pros
  - Exact posterior
  - Easy for on-line learning. *E.g.,*$P(\theta|X) = \beta(N_1+a, N_2+b)$
- Cons
  - In some cases, the conjugate prior may be inadequate

9 Class 10 - Latent Variable

It is a hidden variable that you never observe. *E.g.,*creating a dataset of a job interview, measuring the GPA, IQ, School degree, and Phone, from a first stage of the interview, and a last attribute Onsite performance, from the second interview. Traditional ML models will suffer with the missing data. Even more, probably there are some correlation between the variables, and ponder each combination will increase the number of attributes. Thus, we can link all these attributes with a latent variable, called in this example as Intelligence. We can model the problem as:

$$P(X_1, X_2, X_3, X_4, X_5) = \sum_{i=1}^{N} P(X_1, X_2, X_3, X_4, X_5|I)P(I) \tag{48}$$

where I is the latent variable. We can also simplify the problem with

$$P(X_1, X_2, X_3, X_4, X_5) = \sum_{i=1}^{N} P(X_1|I)P(X_2|I)P(X_3|I)P(X_4|I)P(X_5|I)P(I) \tag{49}$$

which breaks the table in 5 fewer tables, reduce the model complexity and improves the flexibility of the model.

10 Class 12 - Gaussian Mixture Model (GMM)

It is a model of soft clustering with N gaussian's can be described as

$$P(X|\theta) = \pi_1\mathcal{N}(\mu_1, \Sigma_1) + \pi_2\mathcal{N}(\mu_2, \Sigma_2) + ... + \pi_N\mathcal{N}(\mu_N, \Sigma_N) \tag{50}$$

where the θ weight matrix is

$$\theta = \{\pi_1, \pi_2, ..., \pi_N, \mu_1, \mu_2, ..., \mu_N, \Sigma_1, \Sigma_2, ..., \Sigma_N\} \tag{51}$$

To goal of the train is

$$\max_{\theta} P(X|\theta) = \prod_{i=1}^{N} P(X_1|\theta) = \prod_{i=1}^{N} (pi_1\mathcal{N}(\mu_1, \Sigma_1) + ...) \tag{52}$$

By definition, the covariance matrix $\Sigma_k \succ 0$, once time we need to compute Σ^{-1}. Otherwise, we will have divisions by zero.

11 Class 13 - Training a GMM

Assuming that our data points X was generated by a latent variable t. Thus, the probability of a point to belongs to the class c given the parameters θ is

$$P(t = c|\theta) = \pi_c \tag{53}$$

and

$$P(X|t = c, \theta) = \mathcal{N}(X, \mu_c, \Sigma_c) \tag{54}$$

Marginalizing the latent variable, we have

$$P(X|\theta) = \sum_{c=1}^{N} P(X|t = c, \theta)P(t = c|\theta) \tag{55}$$

which is the summation of the likelihood times the prior, and is exactly the same as the first equation, but ignoring the latent variable.

Training a GMM, if we have hard-labels of the points, it is a easy problem: we just need to compute the mean and standard deviation of each cluster. In the opposite side, if we have the parameters (consequently the gaussians), thus we can estimate the source ($P(t = 1|X, \theta) = \frac{P(X|t=1,\theta)P(t=1|\theta)}{Z}$), which is the joint probability (the likelihood times the prior). The problem of train a GMM is a Chicken and Egg problem:

- Need gaussian parameters (θ) to estimate the source (soft-labels)
- Need sources to estimate the gaussian parameters

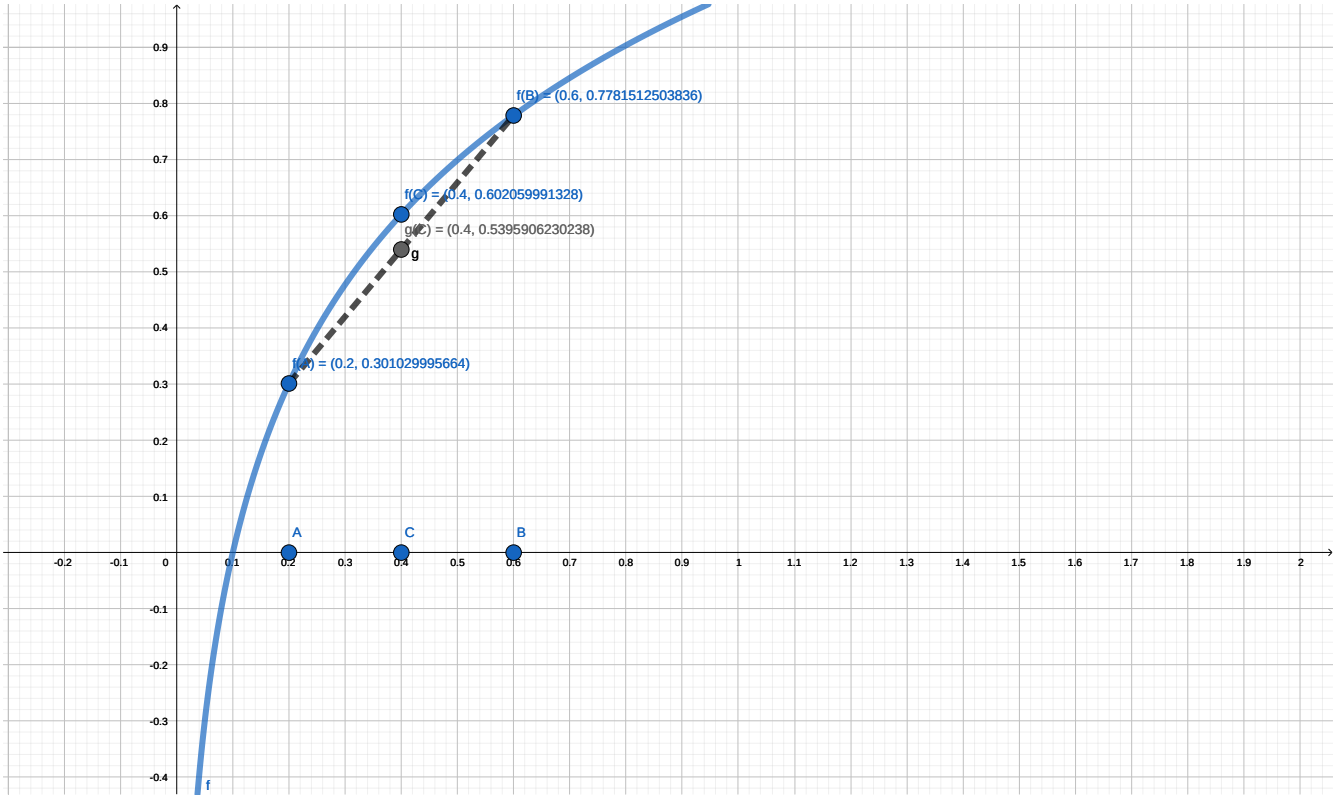


Figure 1: Showing the inequality from the equation. The equation from the right inequality side, which generates the straight segment, we call as G(x).

12 Class 14 - EM GMM

1. Start with N randomly placed gaussians parameters θ
2. Until convergence:
   (a) For each point X_i, compute $P(t = c|X_i, \theta)$
   (b) Update gaussian parameters θ to fit points assigned to them

- EM can train GMM faster than Stochastic Gradient Descent
- EM suffers from local maxima (the exact solution is NP-Hard)

13 Class 15

13.1 Concavity

A function is concave if its second derivative is negative, or, if

$$f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha)f(b) \tag{56}$$

where $0 \leq \alpha \leq 1$. The Figure 1 shows an *e.g.,*

13.2 Jensen's inequality

Generalizing the concavity for any point, we have

$$f(\mathbb{E}_{p(t)t}) \geq \mathbb{E}_{p(t)} f(t) \tag{57}$$

13.3 Kullback-Leibler Divergence

It is a way to measure the difference between two probabilistic functions.

$$\mathcal{KL}(p||q) = \int q(x) \log \frac{q(x)}{p(x)} dx \tag{58}$$

Properties:

1.
$$\mathcal{KL}(p||q) \neq \mathcal{KL}(q||p) \tag{59}$$

2.
$$\mathcal{KL}(q||q) = 0 \tag{60}$$

3.
$$\mathcal{KL}(p||q) > 1 \tag{61}$$

Proof

$$-\mathcal{KL}(p||q) = \mathbb{E}_q \left(-\log \frac{q}{p} \right) = \mathbb{E}_q \left(\log \frac{p}{q} \right) \tag{62}$$

$$\leq \log \left(\mathbb{E}_q \frac{p}{q} \right) = \log \int q(x) \frac{p(x)}{q(x)} dx = 0 \tag{63}$$

14 Class 16 - Expectation Maximization

Using the log for mathematical conveniences, we want to

$$\max_\theta \log(P(X|\theta)) = \log \left(\prod_{i=1}^{N} p(x_i|\theta) \right) \tag{64}$$

with the log properties, we have

$$\max_\theta \log(P(X|\theta)) = \log \left(\sum_{i=1}^{N} \log(p(x_i|\theta)) \right) \tag{65}$$

The probability $\log(p(x|\theta))$ is

$$\log(p(x|\theta)) = \log \left(\sum_{i=1}^{N} \log(p(x_i|\theta)) \right) \tag{66}$$

which we can change the marginal likelihood of the data object x_i by the definition, resulting in

$$= \sum_{i=1}^{N} \log \left(\sum_{c=1}^{C} p(x_i, t_i = c|\theta) \right) \tag{67}$$

with the Jensen's inequality, we have

$$= \sum_{i=1}^{N} \log \left(\sum_{c=1}^{C} p(x_i, t_i = c|\theta) \right) \geq \mathcal{L}(\theta) \tag{68}$$

which means that instead of maximize the original marginal log likelihood, we will maximize a lower bound instead, which is more easy to maximize. Multiplying a dividing a term by the same value, we don't change the function. So, for convenience, we have

$$= \sum_{i=1}^{N} \log \left(\sum_{c=1}^{C} \frac{q(t_i = c)}{q(t_i = c)} p(x_i, t_i = c|\theta) \right) \tag{69}$$

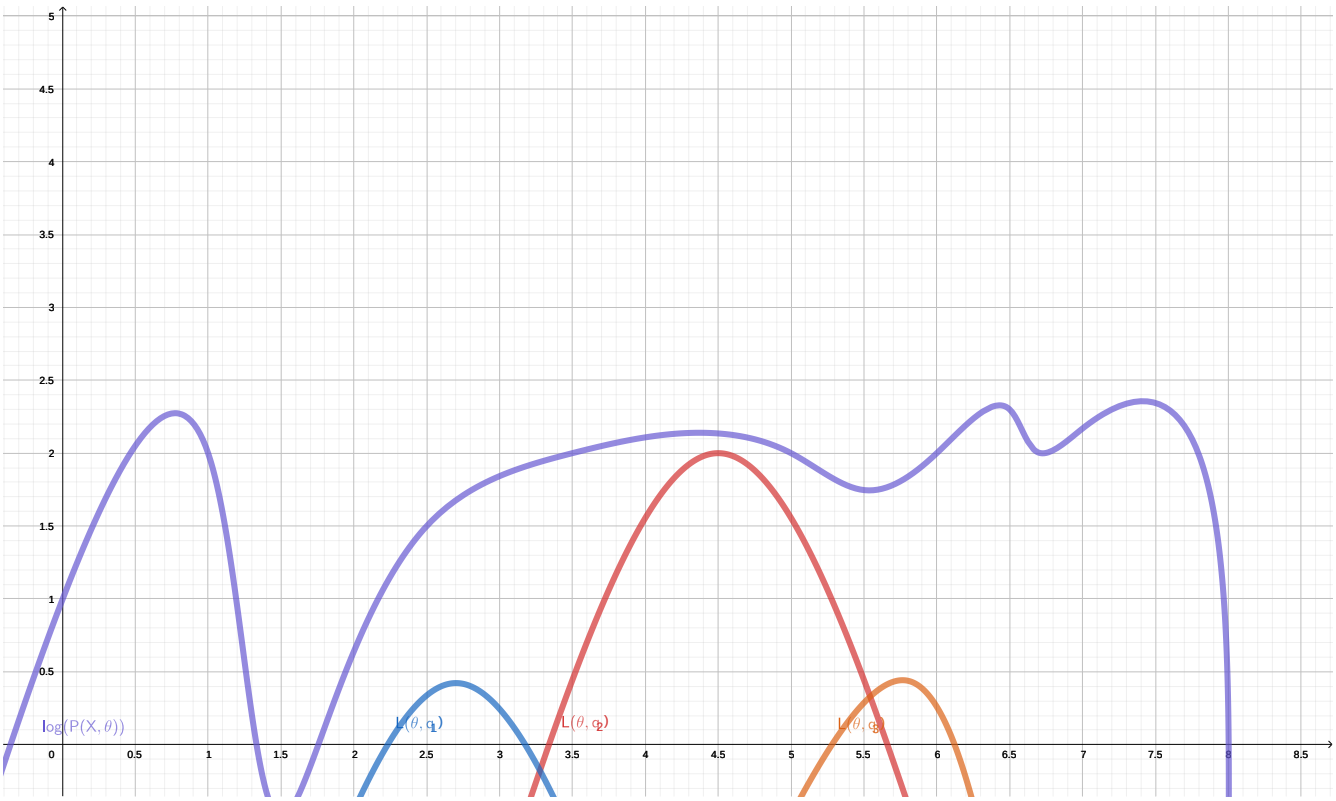


Figure 2: Loss from the General Form of Expectation of Maximization, changing q value. θ is the x-axis.

rewriting, what we have is the Jensen's inequality, in this equation

$$\log\left(\sum_c \alpha_c \upsilon_c\right) = \sum_c \left(\log \alpha_c(\upsilon_c)\right) \tag{70}$$

applying the logarithm properties from the Jensen's inequality, we rebuild the function to

$$\geq \sum_{i=1}^{N}\sum_{c=1}^{C}\left(q(t_i = c)\log\frac{p(x_i, t_i = c|\theta}{q(t_i = c)}\right) \tag{71}$$

$$= \mathcal{L}(\theta, q) \tag{72}$$

Graphically, we are changing the loss as we change the value of q, as in Figure 2. In practice, fixing q we have a new loss $\mathcal{L}$, as the blue curve in figure, which tends to have an global optimum point similar to a local optimum point in the original curve, the purple. In the next q step, we have another curve. Repeating iteratively, these variational lower-bound curves tends to lead θ to the global optimal point from the original curve.

Summarizing:

- $\log p(X|\theta) \geq \mathcal{L}(\theta, q)$ for any q, where $\mathcal{L}(\theta, q)$ is the variational lower bound
- **E-step**

$$q^{k+1} = arg\max_q \mathcal{L}(\theta^k, q) \tag{73}$$

- **M-step**

$$\theta^{k+1} = arg\max_\theta \mathcal{L}(\theta, q^{k+1}) \tag{74}$$

15 Class 17 - E Step

Fixing θ and maximizing $\mathcal{L}$, which is a log likelihood, changing q, which is a distribution

$$\max_q \mathcal{L}(\theta^k, q) \tag{75}$$

as in Figure 2, we want to minimize the gap between the purple curve, the real log likelihood $\log P(X|\theta)$, and some another curve, which is fixed the current lower bound at step k. We can describe the gap as

$$gap = \log P(X|\theta) - \mathcal{L}(\theta^k, q) \tag{76}$$

which is equal to

$$= \sum_{i=1}^{N} \log P(X_i|\theta) - \sum_{i=1}^{N}\sum_{c=1}^{C} q(t_i = c) \log \frac{P(X_i, t_i = c|\theta)}{q(t_i = c)} \tag{77}$$

rearranging the summations, we have

$$= \sum_{i=1}^{N} \left(\log P(X_i|\theta) * \sum_{c=1}^{N} q(t_i = c) - \sum_{c=1}^{C} q(t_i = c) \log \frac{P(X_i, t_i = c|\theta)}{q(t_i = c)} \right) \tag{78}$$

but $\sum_{c=1}^{N} q(t_i = c)$ is the summation of the probabilities of all the classes, which is always 1. As the two inner summations have this element, we can rearrange the equation, giving

$$= \sum_{i=1}^{N}\sum_{c=1}^{N} q(t_i = c) \left(\log P(X_i|\theta) - \log \frac{P(X_i, t_i = c|\theta)}{q(t_i = c)} \right) \tag{79}$$

and using the logarithm properties of division of terms, we have

$$= \sum_{i=1}^{N}\sum_{c=1}^{N} q(t_i = c) \left(\log \frac{P(X_i|\theta) q(t_i = c)}{P(X_i, t_i = c|\theta)} \right) \tag{80}$$

by some bayesian rules, $P(X_i, t_i = c|\theta) = P(t_i = c|X_i, \theta)P(X_i|\theta)$, replacing the terms, we can simplify the $P(X_i|\theta)$ in the numerator and denominator, giving

$$= \sum_{i=1}^{N}\sum_{c=1}^{N} q(t_i = c) \left(\log \frac{q(t_i = c)}{P(t_i = c|X_i, \theta)} \right) \tag{81}$$

where the inner summation is exactly the Kullback-Leibler Divergence $\mathcal{KL}\left(q(t_i||P(t_i|X_i, \theta)\right)$. Our final equation is

$$gap = \sum_{i=1}^{N} \mathcal{KL}\left(q(t_i||P(t_i|X_i, \theta)\right) \tag{82}$$

As we want to maximize the lower bound $\mathcal{L}$, we want to minimize the difference gived by the first gap equation, which means minimize the summation of the Kullback-Leibler Divergence.

16 Class 18 - M Step

We want to maximize the lower-bound log likelihood, as discussed in class 16 after rebuild the equation in a Jensen's inequality, with the following equation

$$\mathcal{L}(\theta, q) = \sum_{i=1}^{N}\sum_{c=1}^{N} q(t_i = c) log \frac{P(X_i, t_i = c|\theta)}{q(t_i = c)} \tag{83}$$

but now, in terms of θ. With the division logarithm property, we can transform to the following equation to

$$\mathcal{L}(\theta, q) = \sum_{i=1}^{N}\sum_{c=1}^{N} q(t_i = c) log P(X_i, t_i = c|\theta) - \sum_{i=1}^{N}\sum_{c=1}^{N} q(t_i = c) log q(t_i = c) \tag{84}$$

but, as we can maximize in terms of θ, and the second term haven't θ, we can throw this term away, which results in

$$\mathbb{E}_q \log P(X, T|\theta) + const \tag{85}$$

which usually is a concave function, so it is easy to maximize. The goal in the M-step is

16.1 Summarising Expectation Maximization

E-step

$$q^{k+1} = arg\min_{q} \mathcal{KL}[q(T)||P(T|X,\theta^k)] \Leftrightarrow q^{k+1}(t_i) = p(t_i|x_i,\theta^k) \tag{86}$$

M-step

$$\theta^{k+1} = arg\max_{\theta} \mathbb{E}_{q^{k+1}} \log P(X,T|\theta^k) \tag{87}$$

17 Class 19 - E step example

Imagine a dataset with three values: $1, 2$, and 3, with $N_1 = 30, N_2 = 20$, and $N_3 = 60$, respectively. Threre are two gaussians to fit the data, parametrized by the parameters α, β, and γ. The probability function is $P(x_i) = \gamma P_1(x_i) + (1-\gamma)P_2(x_i)$. The following table gives the probability distribution to each value accordingly to each P_n: where $\alpha_0 = \beta_0 = \gamma_0 = 0.5$. Using the latent variable t we have

	1	2	3
P_1	α	$1-\alpha$	0
P_1	0	$1-\beta$	β

Table 1: Probabilities to sample the values $1, 2$, and 3 with the probability models P_1 and P_2.

$$P(t_i = 1) = \gamma \tag{88}$$

$$P(x_i|t_i = 2) = P_2(x_i) \tag{89}$$

By definition, $q(t_i) = P(t_i = c|x_i)$, and $P(X|Y) = \frac{P(Y|X)P(Y)}{P(X)}$. Changing the terms of this equation by the values given by the example, we have

$$q(t_i = 1) = P(t_i = 1|x_i = 1) = \frac{P(x_i = 1|t_i = 1)P(t_i = 1)}{P(x_i = 1|t_i = 1)P(t_i = 1) + P(x_i = 2|t_i = 2)P(t_i = 2)} \tag{90}$$

looking for Table 1, we have

$$= \frac{\alpha\gamma}{\alpha\gamma + 0(1-\alpha)} = 1 \tag{91}$$

and for $t = 2$

$$= \frac{(1-\alpha)\gamma}{(1-\alpha)\gamma + (1-\beta)(1-\gamma)} = \frac{0.5 * 0.5}{0.5 * 0.5 + 0.5 * 0.5} = 0.5 \tag{92}$$

18 Class 20 - M step example

Following the example, we have $N_1 = 30, N_2 = 20$, $\alpha_0 = \beta_0 = \gamma_0 = 0.5$, $P(x_i) = \gamma P_1(x_i) + (1-\gamma)P_2(x_i)$, the Table 1 with the probabilities, and we discover in E step that

$$q(t_i = 1) = P(t_i = 1|x_i) = \begin{cases} 1, & x_i = 1 \\ 0.5, & x_i = 2 \\ 0, & x_i = 3 \end{cases} \tag{93}$$

and $q(t_i = 2) = 1 - q(t_i = 1)$. By definition, we want a maximization with

$$\max_{\alpha,\beta,\gamma} \sum_{i=1}^{N} \mathbb{E}_{q(t_i)} \log p(x_i|t_i)p(t_i) = \sum_{i=1}^{N} q(t_i = 1) \log(p(x_i|t_i = 1))\gamma + \sum_{i=1}^{N} q(t_i = 2) \log(p_2(x_i))(1-\gamma) \tag{94}$$

Changing the constants by the values, we have

$$= 30P(t_i = 1|x_i = 1) \log(\alpha)\gamma + 20 * 0.5 \log(1-\alpha)\gamma + 60 * 0 * \log 0 + 20 * 0.5 \log(1-\beta)(1-\gamma) + 60 * 1 * \log(\beta)(1-\gamma) \tag{95}$$

as $P(t_i = 1|x_i = 1) = 1$ and other terms zeroed, we have

$$30 \log \alpha + 1 \log(1-\alpha) + const(\alpha) \tag{96}$$

which is the problem that we want to maximize. Using the gradient (deriving the logs), we have

$$30\frac{1}{\alpha} + 10\frac{1*(-1)}{1-\alpha} = 0 \rightarrow \frac{30}{\alpha} = \frac{10}{1-\alpha} \rightarrow 30 - 30\alpha = 10\alpha \rightarrow 40\alpha = 30 \rightarrow \alpha = \frac{3}{4} = 0.75 \tag{97}$$

and finally, $\beta = \frac{6}{7}$ and $\gamma = \frac{4}{11}$.

19 Class 22 - General EM for GMM

1. E-step
   - EM: for each point, compute $q(t_i) = p(t_i|x_i, \theta)$
   - GMM: equally, compute $p(t_i|x_i, \theta)$
2. M-step
   - EM: update parameters to maximize $\max_\theta \mathbb{E}_q \log p(X, T|\theta)$
   - GMM: update Gaussian parameters to fit points assigned to them $\mu_i = \frac{\sum_i p(t_i=1|x_i,\theta)x_i}{\sum_i p(t_i=1|x_i,\theta)}$. *E.g.,*

$$\max_\theta \mathbb{E}_q \log p(X, T|\theta) = \sum_{i=1}^{N}\sum_{c=1}^{C} q(t_i = c) \log\left(\frac{1}{z} e^{-\frac{(x_i-\mu_c)^2}{2\sigma_c^2}} \pi_c\right) \tag{98}$$

$$= \sum_{i=1}^{N}\sum_{c=1}^{C} q(t_i = c) \log(\frac{\pi_c}{z}) - -\frac{(x_i - \mu_c)^2}{2\sigma_c^2} \tag{99}$$

and deriving it in respect to the first Gaussian, which uses μ_1, we have

$$\frac{\partial ...}{\partial_{\mu_c}} = \sum_{i=1}^{N} q(t_i = 1)(0 - \frac{2(x_i - \mu_{i=1})(-1)}{2\sigma_{c=1}^2}) \tag{100}$$

multiplying the term by $\sigma_{c=1}^2$ to eliminate the term in the denominator, and setting equal to zero (once time the lower bound log likelihood is a concave curve, and the derivative zero is in the highest point), we have

$$= \sum_{i=1}^{N} q(t_i = 1)x_i - \sum_{i=1}^{N} q(t_i = 1)\mu_{i=1} = 0 \tag{101}$$

$$\mu_1 = \frac{\sum_{i=1}^{N} q(t_i = 1)x_i}{\sum_{i=1}^{N} q(t_i = 1)} \tag{102}$$

and consequently,

$$\sigma_c^2 = \frac{\sum_{i=1}^{N} q(t_i = c)(x_i - \mu_c)}{\sum_{i=1}^{N} q(t_i = c)} \tag{103}$$

where μ_c here is the new μ computed by the previous one equation. For the gaussian ponder π, we have to ensure that $\pi_c > 0$ and $\sum_{c=1}^{C} \pi_c = 1$. We can update π with the following equation

$$\pi_c = \frac{\sum_{i=1}^{N} q(t_i = c)}{N} \tag{104}$$

20 Class 23

20.1 K-Means from GMM Perspective

In GMM, if we fix the covariances to be identical, $\Sigma_c = \mathcal{I}$, and fix the weights of each gaussian to be uniform, $\pi_c = \frac{1}{NGaussians}$, thus, the conditional probability is

$$p(x_i|t_i = c, \theta) = \frac{1}{Z} e^{-0.5||x_i - \mu_c||^2} \tag{105}$$

which is a weighted euclidean distance between the data and the centroid. Thus, we prove that the K-Means is a special case of the GMM.

20.2 K-Means from EM perspective

The $q(t)$ is approximated by a delta function, which binaries the probabilities

$$q^{k+1}(t_i) = \begin{cases} 1, & t_i = c_i \\ 0, & otherwise \end{cases} \tag{106}$$

then, the **E-step** turns

$$c_i = arg\max_c p(t_i = c|x_i, \theta) \tag{107}$$

where the probability $p(t_i = c|x_i, \theta)$ is

$$p(t_i = c|x_i, \theta) = \frac{1}{Z} p(x_i|t_i, \theta) p(t_i\theta) \tag{108}$$

$$= \frac{1}{Z} e^{-0.5||x_i - \mu_c||^2} \pi_c \tag{109}$$

as $\frac{1}{Z}$ and π_c does not depend on c, we can throw it away. Then, the **E-step** becomes to

$$c_i = arg\max_c p(t_i = c|x_i, \theta) = arg\min_c ||x_i - \mu_c||^2 \tag{110}$$

which is the Euclidean distance used in K-Means.

21 Class 24 - K-Means from EM perspective

The **M-step** is

$$\max_\mu \sum_{i=1}^{N} \mathbb{E}_{q(t_i)} \log(p(x_i, t_i|\mu)) \tag{111}$$

where μ is relative to the parameters from the original equation. As $q(t_i)$ is restricted, as the equation bellow,

$$q(t_i) = \begin{cases} 1, & t_i = c_i^* \\ 0, & t_i \neq c_i^* \end{cases} \tag{112}$$

and do not contribute in the equations, we can rewrite the μ update to

$$\mu_c = \frac{\sum_{i=1}^{N} q(t_i = c) x_i}{\sum_{i=1}^{N} q(t_i = c)} = \frac{\sum_{i:c_i^* = c_i}^{N} x_i}{Ni : c_i^* = c_i} \tag{113}$$

which is exactly the mean of the classified points, as the K-Means do.

22 Class 25 - Probabilistic Principal Component Analysis (PPCA)

We can model a PCA as a probabilistic model using latent variables. Thus, we can threat missing data. Imaging a PCA reducing two dimensions to the dimension t. Thus, threating as a latent variable, we have

$$p(t_i) = \mathcal{N}(0, I) \tag{114}$$

and the points that precisely overlaps the PCA curve are computed by

$$x_i = \mathcal{W}t_i + b \tag{115}$$

while the original data from the dataset can be estimated as

$$x_i = \mathcal{W}t_i + b + \epsilon_i \tag{116}$$

where σ_i is a random noise with ϵ_i $\mathcal{N}(0, \Sigma.$ probabilistically, we have

$$p(x_i|t_i, \theta) = \mathcal{N}(\mathcal{W}t_i + b, \Sigma) \tag{117}$$

As we have continuous variables, we need the Marginalization to train the model

$$\max_{\theta} p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N} \int p(x_i|t_i,\theta)p(t_i)dt_i \tag{118}$$

which is intractable, as the integral means summing all the points. Applying a conjugacy in the integral, resulting in $\mathcal{N}(\mu_i, \Sigma_i)$, we can treat this problem training with Expectation Maximization.

23 Class 26 - EM for PPCA

E-step

$$q(t_i) = p(t_i|x_i,\theta) = \frac{p(x_i|t_i,\theta)p(t_i)}{Z} = \mathcal{N}(\hat{\mu_i}), \hat{\Sigma}_i \tag{119}$$

M-step

$$\max_{\theta} \mathbb{E}_{q(T)} \sum_i \log\left(p(x_i|t_i,\theta)p(t_i)\right) = \sum_i \mathbb{E}_{q(t_i)} \log\left(\frac{1}{Z} e^{\cdots} e^{\cdots}\right) \tag{120}$$

as $\frac{1}{Z}$ not depends of t_i, and with the logarithm property of log multiplication is equal to the sum of the logs, we rewrite to

$$= \sum_i \log(\frac{1}{Z}) + \sum_i \mathbb{E}_{q(t_i)} \log\left(e^{\cdots} e^{\cdots}\right) \tag{121}$$

$$= \sum_i \log(\frac{1}{Z}) + \sum_i \mathbb{E}_{q(t_i)} \log\left(exp\left(-\frac{(x - \mathcal{W}t_i - b)^2}{2\sigma^2}\right) exp\left(-\frac{t_i^2}{2}\right)\right) \tag{122}$$

where all the log term is similar to $at_i^2 + ct_i + d$.

24 Class 27 - Inference Approximation

In some cases, the we can approximate a distribution by another, as the marginal probability can be hard to compute. We can, *e.g.,*, train a neural network to learn a distribution, switching the marginal probability by a constant normalization

$$p^*(z) = p(z|X) = \frac{p(X|z)p(Z)}{p(X)} = \frac{\hat{p}(Z)}{Z} \tag{123}$$

In this manner, we need to minimize the $\mathcal{KL}$-Divergence

$$\mathcal{KL}[q(z)||\frac{\hat{p}(Z)}{Z}] = \int q(z) \log\left(\frac{q(z)}{\hat{p}(Z)/Z}\right) dz = \int q(z) \log\left(\frac{q(z)}{\hat{p}(Z)}\right) dz + \int q(z) \log Z dz \tag{124}$$

$$\mathcal{KL}[q(z)||\hat{p}(Z)] + \log Z \tag{125}$$

as $\log Z$ is a constant, we have

$$\mathcal{KL}[q(z)||\hat{p}(Z)] \to \min_z \tag{126}$$

25 Class 28 - Mean Field

1. Select a family of distributions $\mathcal{Q}$ as a variational inference

$$\mathcal{Q} = q|q(z) = \prod_{i=1}^{d} q_i(z) \tag{127}$$

   which is a set of all distributions that have factorized over the dimensions of the latent variables.

2. optimize to find the best approximation $q(Z)$ of $\hat{p}(Z)$

$$\mathcal{KL}[q(z)||\hat{p}(Z)] \to \min_{q \in \mathcal{Q}} \tag{128}$$

E.g.,, factorizing a distribution of two variables to a multiplication of two distributions: $p^*(z_1, z_2) \approx q_1(z_1)q_2(z_2)$, where p^* is the real distribution. Imaging that $p^*(z_1, z_2) = \mathcal{N}(0, \Sigma)$. Thus, we have $q_1(z_1)q_2(z_2) = \mathcal{N}(0, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix})$. Then, optimize the $\mathcal{KL}$-Divergence for each distribution,

$$\mathcal{KL}(\prod_i^d q_i || p^*) \rightarrow \min_q \tag{129}$$

and the $\mathcal{KL}$-Divergence is

$$= \int \prod_i^d \log\left(\frac{\prod_i^d q_i}{p^*}\right) dz = \sum_i^d \int \prod_j^d \log(q_i) dz - \int \prod_j^d \log(p^*) dz \tag{130}$$

where, for a specific distribution k, we eliminate the first summation and generate

$$= \int \prod_j^d \log(q_k) dz + \sum_{i \neq k}^d \int \prod_j^d \log(q_i) dz - \int \prod_j^d \log(p^*) dz \tag{131}$$

but, some therms are constants in relation to $q(z)$ and its respective integrals are equal to 1. *E.g.,*, the first integral can be decomposed to

$$\int \prod_j^d \log(q_k) dz = \int q_k \log(q_k) \left[\int \prod_{j \neq k}^d \log(q_j) dz_{\neq k} \right] dz_k = \int q_k \log(q_k) dz_k \tag{132}$$

The second integral not depends on k, and thus it is constant. Rewriting, we have

$$\int \prod_i^d \log\left(\frac{\prod_i^d q_i}{p^*}\right) dz = \int q_k \log(q_k) dz_k - \int q_k \left[\prod_{j \neq k}^d q_j \log(p^*) dz_{\neq k} \right] dz_k \tag{133}$$

grouping the similar term q_k, we have

$$= \int q_k \left[\log(q_k) dz_k - \int \prod_{j \neq k}^d q_j \log(p^*) dz_{\neq k} \right] dz_k \tag{134}$$

and the term with the productory is the expectation, $h(z_k) = \mathbb{E}_{q_{-k}} \log p^* = prod_{j \neq k}^d q_j \log(p^*) dz_{\neq k}$. If we rewrite $h(z)$ as $t(z_k) = \frac{e^{h(z_k)}}{\int e^{h(z_k)} dz_k}$, we can rewrite the last equation to a $\mathcal{KL}$-Divergence

$$= \int q_k \log\left(\frac{q_k}{t}\right) dz_k + const = \mathcal{KL}(q_k || t) \rightarrow \min \tag{135}$$

as we want to zero the $\mathcal{KL}$-Divergence, then

$$q_k = t \rightarrow \log q_k = \mathbb{E}_{q_{-k}} \log p^* + const \tag{136}$$

26 Class 30 - Variational EM Review

In EM, we want to maximize the $\mathcal{L}(\theta, q)$, which is equal to $\mathbb{E}_{q(T)} \log \frac{p(X,T|\theta)}{q(T)}$. The $\mathcal{L}(\theta, q)$ is the variational lower bound of the original marginal likelihood $\log p(X|\theta)$. In **E-step**, we want to minimize the $\mathcal{KL}$-Divergence, equaling $q(T)$ to the Full Posterior $p(T|X, \theta)$. The EM becomes to Variational EM if we want to minimize $q \in \mathcal{Q}$. Based on the problem, we can perform different inferences, described below. The first one approaches are more accurately and slow, while the latter ones are less precise but more faster.

- Full inference: $p(T, \theta|X)$ is the precise inference, however, sometimes it is not possible.
- Mean field: $q(T)q(\theta) \approx p(T, \theta|X)$
- EM Algorithm: $q(T), \theta = \theta_{MP}$ estimates only one point of θ
- Variational EM: $q_1(T_1)...q_d(T_d), \theta = \theta_{MP}$ factorizes the latent variables for each dimension
- Crisp EM: $T = T_{MP}, \theta = \theta_{MP}$ with estimate the latent variable and the parameter with point estimate. It is similar to a K-Means estimation.

27 Class 32 - Dirichlet Distribution

It is a distribution which can assume different shapes

$$Dir(\theta, \alpha) = \frac{1}{\beta(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{137}$$

as α changes. This distribution have the following attributes:

- $\mathbb{E}\theta_i = \frac{\alpha_i}{\alpha_0}$
- $Cov(\theta_i, \theta_j) = \frac{\alpha_i \alpha_0 [i=j] - \alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}$
- $\alpha_0 = \sum_{k=1}^{K} \alpha_k$

Looking for this properties, we can conclude that the β distribution is a special case of the Dirichlet Distribution with only 2 dimensions.

The Dirichlet Distribution prior is conjugate to the Multinomial Likelihood. Remembering, the prior ($P(\theta)$) is conjugate to the likelihood ($P(X|\theta)$) if the prior and the posterior ($P(\theta|X)$) are both in the same family of distribution. The Multinomial Likelihood is given by

$$P(X|\theta) = \frac{n!}{x_1!...x_k!} \theta_1^{x_1} ... \theta_k^{x_k} \tag{138}$$

while the Dirichlet prior $p(\theta)$ was gived by the previous equation. Thus, the posterior probability is

$$p(\theta|X) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k + x_k - 1} \tag{139}$$

which is a β distribution with a normalization constant. Thus, we have

$$p(\theta|X) = Dir\left(\theta | \begin{pmatrix} ... \\ \alpha_k + x_k \\ ... \end{pmatrix}\right) \tag{140}$$

28 Class 33 - Latent Dirichlet Allocation (LDA) Model

We can use the Dirichlet distribution to select words in a topic based problem. For instante, a document is a set of topics, where each topics is a set of words, each word with some probability in the correspondent topic. If we create a model Θ, as a Bayesian Network, to select topics as a latent variable z, which impacts in the selection of the words w. Thus, we have

$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}) \tag{141}$$

where

- $\prod_{d=1}^{D}$ to iterate over each document from the dataset.
- $p(\theta_d)$ to generate the d-th topic probability. We can compute this probability by $p(\theta_d \approx Dir(\theta_d, \alpha)$.
- $p(z_{dn}|\theta_d)$ to select a topic based on the latent variable. We can compute this probability as $p(z_{dn}|\theta_d) = \theta_{dz_{dn}}$.
- $p(w_{dn}|z_{dn})$ to select a word based on the topic. We can compute this probability with $p(w_{dn}|z_{dn}) = \Phi_{z_{dn}w_{dn}}$, which is a table with the probabilities.

Thus, this problem contains

- Known W data
- Unknown Φ parameters, with the distribution over words for each topic
- Unknown Z latent variables, topic of each word
- Unknown Θ latent variables, a distribution over topics for each document

29 Class 36 - LDA M-step

To train the LDA with EM, we have the two steps:

- **E-step**, with

$$\mathcal{KL}(q(\theta)q(Z)||p(\theta,Z|W)) \rightarrow \min_{q(\theta),q(Z)} \tag{142}$$

- **M-step**, with

$$\mathbb{E}_{q(\theta)q(Z)} \log p(\theta,Z,W) \rightarrow \max_{\Phi} \tag{143}$$

where the probability $p(\theta,Z,W)$ is given by

$$p(\theta,Z,W) = \sum_{d=1}^{D}\left[\sum_{t=1}^{T}(\alpha_t-1)\log\theta_{dt} + \sum_{n=1}^{N_d}\sum_{t=1}^{T}[z_{dn}=t](\log\theta_{dt}\log\Phi_{tw_{dn}})\right] \tag{144}$$

as the most of the terms has no Φ, thus is constant, we can rewrite the expectation to

$$\mathbb{E}_{q(\theta)q(Z)} \log p(\theta,Z,W) = \mathbb{E}_{q(\theta)q(Z)} \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T}[z_{dn}=t]\log\Phi_{tw_{dn}} + const \rightarrow \max_{\Phi} \tag{145}$$

As Φ is a probability, we need to ensure two constraints: that $\Phi_{tw} \geq 0$, which is always true due to the logarithm; and $\sum \Phi_{tw} = 1$. To ensure this second constraint, we need to use the Lagrangian Multiplier, which is a strategy for finding the local maxima and minima of a function subject to equality constraints. This Lagrangian is given by

$$\mathbb{L} = \mathbb{E}_{q(\theta)q(Z)} \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T}[z_{dn}=t]\log\Phi_{tw_{dn}} + \sum_{t=1}^{T}\lambda_t(\sum_{w}\Phi_{tw}-1) \tag{146}$$

Simplifying this equation, we have

$$= \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T}\gamma_{dn}^{t}\log\Phi_{tw_{dn}} + \sum_{t=1}^{T}\lambda_t(\sum_{w}\Phi_{tw}-1) \tag{147}$$

To maximize, we use the derivatives equalized to zero, given by

$$\frac{\partial\mathbb{L}}{\Phi_{tw}} = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T}[w_{dn}=w]\gamma_{dn}^{t}\frac{1}{tw_{dn}} + \lambda_t = 0 \tag{148}$$

and thus, we have that Φ is

$$\Phi_{tw} = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T}[w_{dn}=w]\gamma_{dn}^{t}}{-\lambda_t} \tag{149}$$

which even depends of λ. As the summation of all the w's is one, by our first constraint, thus

$$\sum_{w}\Phi_{tw} = \sum_{w}\frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T}[w_{dn}=w]\gamma_{dn}^{t}}{-\lambda_t} = 1 \tag{150}$$

isolating λ, we have

$$\lambda_t = \sum_{w,d,n,t}[w_{dn}=w]\gamma_{dn}^{t} \tag{151}$$

Plugging γ in the Φ equation, we have

$$\Phi_{tw} = \frac{\sum_{d,n,t}[w_{dn}=w]\gamma_{dn}^{t}}{\sum_{w,d,n,t}[w_{dn}=w]\gamma_{dn}^{t}} \tag{152}$$

30 Class 37 - Extensions of LDA

Interpreting the LDA:

- $p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn})$
- $p(\theta_d) \approx Dir(\alpha)$
- if α is higher, means that are more topics for each document
- if α is smaller, means that are less topics for each document
- α can be selected as $p(W|\alpha) \rightarrow \alpha$

If we have a sparse prior on Φ, we can work around this problem with

$$p(W, Z, \Theta, \Phi) = \prod_{t=1}^{T} p(\Phi_t)p(W, Z, \Theta|\Phi) \tag{153}$$

where $p(\Phi_t) \approx Dir(\beta)$. We also can compute a **Topic Correlation** as

$$p(\theta_d) \approx P(\mathcal{N}(\mu, \Sigma)) \tag{154}$$

And finally, we can use LDA as a **Dynamic Topic Model**. *E.g.,*where some word is very used at some time and it use decreases along the time, measured as τ. We can do this with the equations bellow

$$p(B_{t\bullet}^{\tau+1}|B_{t\bullet}^{\tau}) \approx \mathcal{N}(B_{t\bullet}^{\tau}, \sigma^2\mathcal{I}) \tag{155}$$

and

$$\Phi_{t\bullet}^{\tau+1} = Softmax[B_{t\bullet}^{\tau}] \tag{156}$$

31 Class 39 - Sampling from 1D distributions

- Discrete Distributions: each category have its own probability, which all summed is one. We can sampling by, generating a number $r \approx \mathcal{U}[0, 1]$ and choosing the category where the sampled value falls inside the range of respective category.
- Continuous Distributions: we cam sample, *e.g.,*using the Central Limit Theorem. Exemplifying, if we have a Gaussian $\mathcal{N}(0, 1)$, we can compute

  $$z = \sum_{i=1}^{12} x_i - 6, x_i \approx \mathcal{U}[0, 1] \tag{157}$$

  where 12 is a chosen number of points and 6 if the half of the number of points. Thus, $p(z) \approx \mathcal{N}(0, 1)$.
- Continuous Unknown Distribution: with we have a real and unknown distribution $p(x)$, we first choose a upper-bound distribution $q(x)$ where $p(x) \leq 2q(x)$. Thus, we need to sample the points $x \approx q(x)$ and $y \approx \mathcal{U}[0, 2q(x)]$. Then, we accept x as a point sampled from $p()$ with probability $\frac{p(x)}{2q(x)}$, if $y \leq p(x)$. If we want to ensure a more probability of p sampling, we can use a normalization constant M, with $p(x) \leq Mq(x)$, and thus only $\frac{1}{M}$ of points will be acceptable.

32 Class 40 - Markov Chains

Markov Chains are a graph probability representation of a problem. As in a graph, it contains states, with edges linking it. Every edge contains a probability. We can use MC to perform sampling in more complicated probability functions.

*E.g.,*with the MC bellow, in the first step, we start on Start, which move us to R. Thus, in the second step, we have 30% of probability to keep in R and 70% to go to L. What is the probability of stay in state R in the third step?

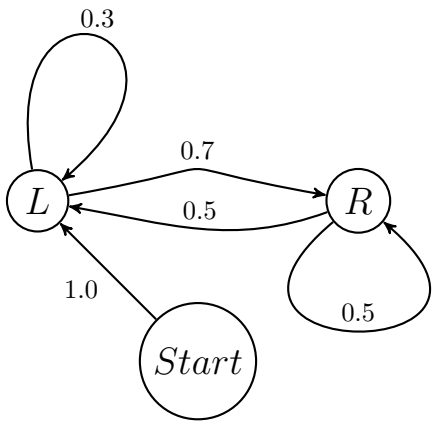


Figure 3: Example of Marcov Chain

This probability is $p(x^3) = p(x^3|x^2 = L)p(x^2 = L) + p(x^3|x^2 = R)p(x^2 = R) = 0.3^2 + 0.7 * 0.5$. In other words, we have the probability of, in step 2 (x^2), we stay in state L and go to R, plus the probability of staying in R in x^2 and keep in R.

We can build a MC to model some distribution. The MC not converge always: it converge if the distribution is stationary - given a distribution π and it respective table of distributions T, the distribution π is stationary if

$$\pi(x') = \sum_x T(x \to x')\pi(x) \tag{158}$$

in other words, it is a probability distribution that remains unchanged in the Markov chain as time progresses. π is invariant by the matrix T.

Another important theorem is that, if $T(x \to x') > 0$ for all x, x', then exists a unique π that satisfies $\pi(x') = \sum_x T(x \to x')\pi(x)$, and thus the markov chain converges to π from any starting point.

33 Class 41 - Gibbs Sampling

This is a Monte Carlo Markov Chain (MCMC) approach for sampling. We start with a random point, *e.g.*,$(x_1^0, x_2^0, x_3^0) = (0, 0, 0)$. The value of this point, in many cases, does not matter. Thus, we need to sample each dimension, as in

$$x_1^1 \sim p(x_1|x_2^0, x_3^0) = \frac{\hat{p}(x1, x_2^0, x_3^0)}{Z_1} \tag{159}$$

we repeat this for the second attribute, $x_2^1 \sim p(x_2|x_1^1, x_3^0)$, and the third attribute, $x_3^1 \sim p(x_3|x_1^1, x_2^1)$.

To proof the Gibbs Sampling, we need to show that

$$p(x', y', z') = \sum_{x,y,z} q(x, y, z \to x', y', z')p(x, y, z) \tag{160}$$

where q is the table with distributions of each markovian state. This proof uses discrete attributes: if the attributes are continuous, we can simply change the summations by integrals, and the proof will be the same. We can change the term of q by its respective computation, giving

$$= \sum_{x,y,z} p(x'|y = y, z = z)p(y'|x = x', z = z)p(z'|x = x', y = y')p(x, y, z) \tag{161}$$

rearranging and grouping some terms, we have

$$= p(z'|x', y') \sum_{y,z} \left(p(x'|y, z)p(y'|x', z) \sum_x p(x, y, z) \right) \tag{162}$$

note that the last summation is equal to $p(y, z)$. Joining the first term of the summation with the term from the last summation, we have the joint distribution. Replacing, we have

$$= p(z'|x', y') \sum_z p(y'|x', z) \sum_y p(x', y, z) \tag{163}$$

the last summation is equal to $p(x', y)$, which give us

$$= p(z'|x', y') \sum_z p(x', y', z) \tag{164}$$

where the last summation is equal to $p(x', y')$. Thus, we have

$$= p(x', y', z') \tag{165}$$

which proof to us that, starting from a point $p(x, y, z)$ and perform the Gibbs Sampling, is equal to sampling from $p(x', y', z')$, which is a stationary point that we can achieve starting from any point.

34 Class 42 - Gibbs Sampling

The given example is very visual and difficult to describe in words. Summarizing, with a point $(0, 0)$, using the Gibbs Sampling to sample from a 2D Gaussian, we sample the coordinate x_2 with the conditional probability of $x_1 = 0$, giving x_2'. Thus, we sample x_1 with the conditional probability of $x_2 = x_2'$. Performing many steps, we have the Gaussian samples.

Summarizing the Gibbs Sampling:

- Pros:
  - Reduce multidimensional sampling to sequence of 1D samplings
  - simple to implement
- Cons:
  - highly correlated samples: many samples similar to each other
  - slow convergence
  - not parallel: we need to sample from each dimension sequentially

35 Class 43 - Metropolis-Hasting

It is another MCMC approach for sampling, which tries to avoid the correlated problem from Gibbs Sampling. In this approach, we need to perform,

- For k = 1, 2, ...
  1. Sample x' from a wrong distribution $\mathcal{Q}(x^k \to x')$
  2. Accept proposal x' with probability $A(x^k \to x')$
  3. Otherwise, $x^{k+1} = x^k$

For this, we need the Markov Chain $\mathcal{Q}$ and the Table Distribution T, computed by

$$T(x \to x') = \mathcal{Q}(x \to x')A(x \to x') \tag{166}$$

for all $x \neq x'$. In the case of not update x, thus we need the following equation

$$T(x \to x) = \mathcal{Q}(x \to x)A(x \to x) + \sum_{x' \neq x} \mathcal{Q}(x \to x')(1 - A(x \to x')) \tag{167}$$

To choose A, we use the policy

$$\pi(x') = \sum_x \pi(x)T(x \to x') \tag{168}$$

which makes use of the Detailed Balance definition. This definition says that, if the markov chain probability of reach a state x' from x is the same as reach x from x', as the equation bellow,

$$\pi(x)T(x \to x') = \pi(x')T(x' \to x) \tag{169}$$

then the probability of $\pi(x')$ is the equation before, $\pi(x') = \sum_x \pi(x)T(x \to x')$. The proof is given bellow

$$\sum_x \pi(x)T(x \to x') = \sum_x \pi(x')T(x' \to x) \tag{170}$$

thus, we can get $\pi(x')$ in the rigth equality side and move it out from the summation, once time is is a constant, as $\pi(x')\sum_x T(x' \to x$. Then, we have a summation of all the probabilities, which is always 1. Thus, we have $\sum_x \pi(x)T(x \to x') = \pi(x')$.

36 Class 44 - Choosing a Critic

In the equation bellow, we have the desired distribution π, which we want to sample from, and Markov Chain $\mathcal{Q}$, which we want to fix

$$\sum_x \pi(x)\mathcal{Q}(x \to x')A(x \to x') = \sum_x \pi(x')\mathcal{Q}(x' \to x)A(x' \to x) \tag{171}$$

where $T(x \to x')$ is equal to $\mathcal{Q}(x \to x')A(x \to x')$ and $T(x' \to x)$ is equal to $\mathcal{Q}(x' \to x)A(x' \to x)$. Isolating the critic A, we have

$$\frac{A(x \to x')}{A(x' \to x)} = \frac{\mathcal{Q}(x' \to x)}{\mathcal{Q}(x \to x')} \tag{172}$$

For simplification, lets $\rho = \frac{\mathcal{Q}(x' \to x)}{\mathcal{Q}(x \to x')}$. If we temporally assume that $\rho < 1$, thus we have $A(x \to x') = \rho$ and $A(x' \to x) = 1$, and also

$$A(x \to x') = \min\left\{1, \frac{\mathcal{Q}(x' \to x)}{\mathcal{Q}(x \to x')}\right\} \tag{173}$$

37 Class 45 - Example of Metropolis Hastings

In the example, sampling from a 1D distribution, the Markov Chain $\mathcal{Q}$ is represented as $\mathcal{Q}(x \to x') = \mathcal{N}(x, 1)$, where $x = 0$, in the example, is the local minimum from the real distribution, which seems like a two-modal distribution. Thus, sampling x', it accepts it with probability $A(x \to x') = \min\left\{1, \frac{\mathcal{Q}(x' \to x)}{\mathcal{Q}(x \to x')}\right\} = \frac{\pi(x')}{\pi(x)}$, where π is the real distribution. Thus, this process is repeated iterativelly, until convergence.

Summarizing the Metropolis Hastings have the property of reject samples applied to Markov Chains, which is:

- Pros:
  - You can choose the family of Markov Chain: *e.g.*,$\mathcal{N}(x, 1)$ or $\mathcal{N}(x, 0.2^2)$. As higher the variance, less correlated the points due to a higher oscillation, but increasing the convergence time.
  - Work with unnormalized densities.
  - Easy to implement.
- Cons:
  - Samples are still correlated.
  - Have to choose among family of Markov Chains.

38 Class 46 - Monte Carlo

Monte Carlo was used in Gibbs Sampling and Metropolis Hastings. Its general formula is given by

$$\mathbb{E}_{p(x)}\frac{1}{M}\sum_{s=1}^{M} f(x_s) = \mathbb{E}_{p(x)}f(x_s) \tag{174}$$

Figure 4: Error comparison between Monte Carlo and Variational Inference

which assumes that $x_s \approx p(x)$, it is an unbiased estimation. Thus, as larger M, better is accuracy. This is different from the Variational Inference

$$\mathbb{E}_{p(x)} f(x) \approx \mathbb{E}_{q(x)} f(x) \tag{175}$$

which assumes that $p(x) \approx q(x)$, it accuracy increases as more samples. We can see a graphical comparison in Figure 4, where the y axis is the error and the x axis the time. Bellow, follows a comparison between the methods and its respective approximation of the probability, from the best at top to the worst at bottom:

- Full inference (or Bayesian): $p(T, \theta|X)$
  - Mean field: $q(T)q(\theta) \approx p(T, \theta|X)$
  - MCMC: $T_s, \Theta_s \sim p(T, \Theta|X)$
- EM algorithm: $q(T), \theta = \theta_{MP}$
  - Variational EM: $q_1(T_1)...q_n(T_n), \theta = \theta_{MP}$
  - MCMC EM: $T_s \sim p(T|\Theta, X), \Theta = \theta_{MP}$

39 Class 47 - MCMC for LDA

We can perform text generation with MCMC. Remembering, in text generation we a document, which is a set of topics, and each topic is a set of words, each with a respective probability.

- Known: W data
- Unknown: Φ latent variables, distribution over words of each topic
- Unknown: Z latent variables, topic of each word
- Unknown: Θ latent variables, distribution over topics for each document

E.g.,, we can sample a word with Gibbs Sampling,

$$p(\Phi, \Theta, Z|W) \sim \{Gibbs\} \tag{176}$$

Thus, with an initial Φ^0, Θ^0, Z^0, we update Φ, dimension by dimension, with $\phi_1^1 \sim p(\phi_1|\phi_2^0, \phi_3^0, ..., \Theta^0, Z^0, W)$, $\phi_2^1 \sim p(\phi_2|\phi_1^1, \phi_3^0, ..., \Theta^0, Z^0, W)$, and etc. The same ideia is repeated for Θ and Z.

40 Class 49 - Scaling Variational Inference

Bayesian inference can be performed efficiently in large datasets using Deep Learning.

40.1 Unbiased Estimation

An estimator is unbiased if its expected value is equals to the thing it estimates. *E.g.,*if we have a distribution, we can sample some points and compute the mean, which is the expectation. Then, we can perform another set of samples, generating another mean. Repeating this process iterativelly, we can create a set of means, and the mean of this set tends to be equal to the real mean of the distribution, if the estimator is unbiased. Thus, we have

$$\mathbb{E}_{p(x)} f(x) \sim \frac{1}{M} \sum_{s=1}^{M} f(x_s) = R \tag{177}$$

where R is the real mean.

41 Final

Look at this links to perform our GMM:

- Sampling from a GMM: `https://stats.stackexchange.com/questions/269205/sampling-from-a-multivariate-gaussian-mixture-model`
- Sklearn GMM doc: `https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html`
- Sklearn train example: `https://scikit-learn.org/0.16/modules/generated/sklearn.mixture.GMM.html`
- Variational GMM: `https://scikit-learn.org/stable/modules/mixture.html`