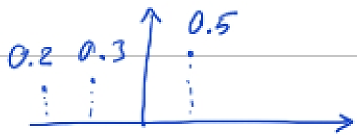
10-Think Bayesian

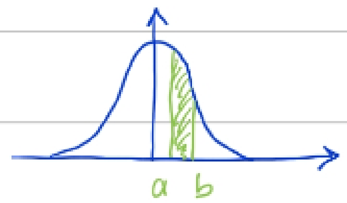
Discrete vs Continuous Distributions



$$P(x) = \begin{cases} 0.2 & x = 1 \\ 0.3 & x = 2 \\ 0.5 & x = 4 \\ 0 & \text{otherwise} \end{cases}$$

$$\sum = 1$$

- In discrete approach $P(x)$ is directly given



$$P(x \in [a,b]) = \int_a^b p(x)\,dx$$

- In continuous approach, the most convenient way is by defining Probability Distribution Function (PDF)

Dependencies

- The two r.v.s variables are considered independent if:

$$\underbrace{P(X,Y)}_{\text{joint}} = \underbrace{P(X)\,P(Y)}_{\text{Marginals}}$$

Conditional Probability

$$\underbrace{P(X|Y)}_{\text{Conditional}} = \frac{\overbrace{P(X,Y)}^{\text{joint}}}{\underbrace{P(Y)}_{\text{Marginal}}}$$

Dependencies

- The two rvn variables are considered independent if:

$$P(X,Y) = P(X)\,P(Y)$$

joint

Marginals

Conditional Probability

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

joint

Marginal

Conditional

Trick 1: Chain Rule

$$P(X,Y) = P(X|Y)\,P(Y)$$

$$P(X,Y,Z) = P(X|Y,Z)\,P(Y|Z)\,P(Z)$$

$$P(X_1,\dots,X_N) = \prod_{i=1}^{N} P(X_i|X_1,\dots,X_{i-1})$$

Note: The intersection (joint probability) is associative/communicative: Therefore:

$$\begin{aligned} P(X,Y,Z) &= P(X|Y,Z)\,P(Y|Z)\,P(Z) \\ &= P(X|Y,Z)\,P(Z|Y)\,P(Y) \\ &= P(Y|X,Z)\,P(Z|X)\,P(X) \\ &\vdots \end{aligned}$$

Trick 2: Sum Rule

Note: The intersection (joint probability) is associative/communicative: Therefore:

$$P(X,Y,Z) = P(X|Y,Z)\,P(Y|Z)\,P(Z)$$
$$= P(X|Y,Z)\,P(Z|Y)\,P(Y)$$
$$= P(Y|X,Z)\,P(Z|X)\,P(X)$$
$$\vdots$$

Trick 2: Sum Rule

$$\underbrace{p(x)}_{\text{marginal distribution}} = \int_{-\infty}^{\infty} \underbrace{p(X,Y)}_{\text{joint probability}}\, d\underbrace{Y}_{\text{random variable}}$$

Bayes Theorem

$$\overbrace{P(\theta|X)}^{\text{posterior}} = \frac{P(\theta,X)}{\underbrace{P(X)}_{\text{evidence}}} = \frac{\overbrace{P(X|\theta)}^{\text{likelihood}}\,\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(X)}_{\text{evidence}}}$$

Frequentist vs Besian

Frequentist	Besian
objective	subjective
considers randomness of events	tries to predict behavior based on evidences
θ is fixed, X is random	θ is random, X is fixed

considers randomness of events

$\begin{cases} \theta \text{ is fixed} \\ X \text{ is random} \end{cases}$

tries to predict behavior based on evidences

$\begin{cases} \theta \text{ is random} \\ X \text{ is fixed} \end{cases}$

works only when the number of data points is much bigger than the number of parameters

$$|X| \gg |\theta|$$

works for arbitrary number of data points

for any $|X|$

For training, it uses Maximum Likelihood

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

For training, it uses Bayes Theorem

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

Training

$$P(\theta|X_{tr}, y_{tr}) = \frac{P(y_{tr}|X_{tr}, \theta)P(\theta)}{P(y_{tr}|X_{tr})}$$

Prediction

$$P(y_{ts}|X_{ts}, X_{tr}, y_{tr}) = \int P(y_{ts}|X_{ts}, \theta)P(\theta|X_{tr}, y_{tr})\,d\theta$$

Regularization

Regularizer

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

online learning

Likelihood

prior

$$P_k(\theta) = P(\theta|x_k) = \frac{P(x_k|\theta)P_{k-1}(\theta)}{P(x_k)}$$

new prior

posterior

Draft Page

-proof of training formula below:

$P(y_{tr}|X_{tr}, \theta)P(\theta)$

-proof of training formula below:

$$* P(\theta|X_{tr}, y_{tr}) = \frac{P(y_{tr}|X_{tr}, \theta)P(\theta)}{P(y_{tr}|X_{tr})}$$

$$P(\theta|X_{tr}, y_{tr}) \underset{\text{Bayes Rule}}{=} \frac{P(\theta, X_{tr}, y_{tr})}{P(X_{tr}, y_{tr})}$$

$$\underset{\text{Chain Rule}}{=} \frac{P(y_{tr}|\theta, X_{tr})P(X_{tr}|\theta)P(\theta)}{P(y_{tr}|X_{tr})P(X_{tr})} = \frac{P(y_{tr}|X_{tr},\theta)P(\theta)}{P(y_{tr}, X_{tr})}$$

Proof of equation of prediction

$$P(y_{ts}|X_{ts}, X_{tr}, y_{tr}) \underset{\text{Sum Rule}}{=} \int P(y_{ts}|X_{ts}, X_{tr}, y_{tr}, \theta)\, d\theta$$

$$P(y_{ts}|X_{ts}, X_{tr}, y_{tr}, \theta) \underset{\text{Bayes Rule}}{=} \frac{P(y_{ts}, X_{ts}, X_{tr}, y_{tr}, \theta)}{P(X_{ts}, X_{tr}, y_{tr}, \theta)}$$

$$=$$

How to define a model?

Bayesian network

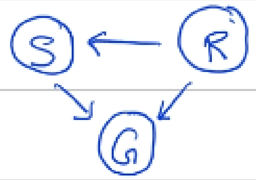
Nodes: random variables

Edges: direct impact

Model: joint probability over all variables

$$P(X_1, \dots, X_n) = \prod_{k=1}^{n} P(X_k \mid \underbrace{Pa(X_k)}_{\text{parents}})$$

ex:

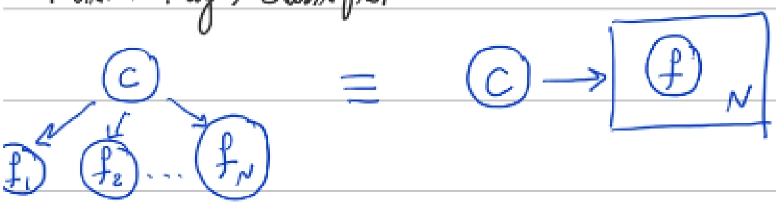


$Pa(G) = \{R, S\}$

$$P(S,R,G) = P(G \mid S,R)\, P(S \mid R)\, P(R)$$

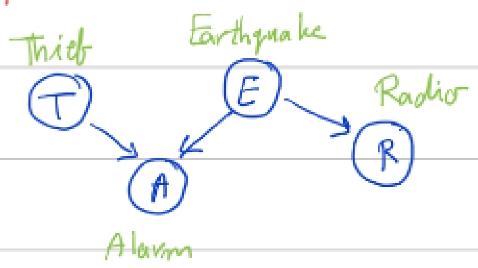
Naïve Bayes classifier

Plate Notation



$$P(c, f_1, f_2, \dots, f_N) = P(c) \prod_{i=1}^{N} P(f_i \mid c)$$

Example : thief & alarm



$$P(t,a,e,r) = ? = P(t)\,P(e)\,P(a|t,e)\,P(r|e)$$

Priors	
$P(T=1)$	10^{-3}
$P(E=1)$	10^{-2}

$P(A=1 \mid T,E)$	$E=0$	$E=1$
$T=0$	0	1/10
$T=1$	1	1

$P(R \mid E)$	
$E=0$	0
$E=1$	1/2

* a notation simplification : $T=1 \longrightarrow T$
$T=0 \longrightarrow \bar{T}$

Sum Rule

$$P(T|A) = \frac{P(T,A)}{P(A)} = \frac{P(T,A,E)+P(T,A,\bar{E})}{P(T,A,E)+P(T,A,\bar{E})+P(\bar{T},A,E)+P(\bar{T},A,\bar{E})}$$

- $P(T,A,E) = P(A|T,E)P(T)P(E) = 10^{-3}\times 10^{-2}\times 1 = 10^{-5}$
- $P(\bar{T},A,\bar{E}) = P(A|\bar{T},\bar{E})P(\bar{T})P(\bar{E}) = 0$
- $P(T,A,\bar{E}) = P(A|T,\bar{E})P(T)P(\bar{E}) = 1\times 10^{-3}\times 0.99 = 9.9\times 10^{-4}$

⋮

$$P(T|A) \simeq 50\%$$

$P(T|A,R) = ?$ what is the probability of a Thief in a house if you hear an alarm & radio report

$$P(T|A,R) = \frac{P(T,A,R)}{P(A,R)} = \frac{P(T,A,R,E) + P(T,A,R,\bar{E})}{P(A,R,T,E) + P(A,R,T,\bar{E}) + P(A,R,\bar{T},E) + P(A,R,\bar{T},\bar{E})}$$

* The earthquake parameter 'E' is added because the event depends on 'E' as a parent. *

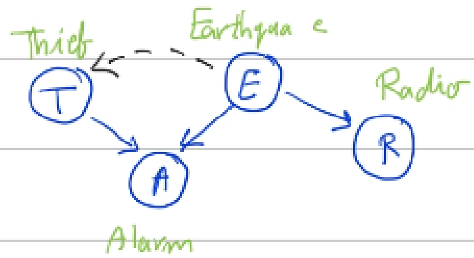
- $P(T,A,R,E) = P(A|T,E)\,P(R|E)\,P(E)$
- $P(T,A,R,\bar{E}) = P(A|T,\bar{E})\,P(R|\bar{E})\,P(\bar{E}) = 1 \times 0 \times 0.99 = 0$

⋮

$$\Rightarrow P(T|A,R) \simeq 1\%$$

* If the results do not match with expectations, the model should be modified. For example:

– There are more Thieves in an event of an earthquake



LINEAR REGRESSION

Univariate Normal

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

Multivariate Normal

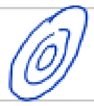
$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

$$\mathbb{E}X = \mu \qquad \mathrm{Cov}[X] = \Sigma$$

mean vector — Covariant Matrix

Σ matrix representations

$$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

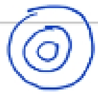


Full

params: $\frac{D(D+1)}{2}$

Diagonal

params: D



Spherical

params: 1

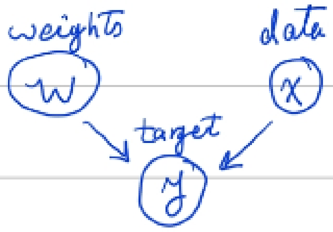
approximation — approximation

Linear Regression

$$L(\omega) = \sum_{i=1}^{N}(\omega^T x_i - y_i)^2 = \|\omega^T X - y\|^2 \longrightarrow \min_{\omega}$$

$$\hat{\omega} = \arg\min_{\omega} L(\omega)$$

∘ Bayesian Model



$$P(\omega, y|X) = P(y|X,\omega)P(\omega)$$

$$P(y|X,\omega) = \mathcal{N}(y|\omega^T X, \sigma^2 I)$$: assuming normal distribution

$P(\omega, y|X) = P(y|X,\omega)\,P(\omega)$

$P(y|X,\omega) = \mathcal{N}(y|\omega^T X, \sigma^2 I)$: assuming normal distribution

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

- Calculate the posterior probability

$$P(\omega|y,X) = \frac{P(y,\omega|X)}{P(y|X)} \longrightarrow \max_{\omega}$$

$P(y,\omega|X)$ → depends on ω

$P(y|X)$ → independent of ω

$$\hookrightarrow P(y,\omega|X) = P(y|X,\omega)\,P(\omega) \longrightarrow \max_{\omega} \longrightarrow \text{take log}$$

$$\log P(y|X,\omega) + \log P(\omega) =$$

$$\log C_1 \exp\left(-\tfrac{1}{2}(y-W^TX)^T[\sigma^2 I]^{-1}(y-W^TX)\right) +$$

$$\log C_2 \exp\left(-\tfrac{1}{2}\omega^T[\gamma^2 I]^{-1}\omega^T\right)$$

$$= \frac{-1}{2\sigma^2}\underbrace{(y-\omega^TX)^T(y-W^TX)}_{\|y-\omega^TX\|^2} - \frac{1}{2\gamma^2}\underbrace{\omega^T\omega}_{\|\omega\|^2} \longrightarrow \max_{\omega} \quad \Big| \times -1 \times 2\sigma^2$$

$$\Rightarrow \underbrace{\|y-\omega^TX\|^2}_{\text{sum of squares}} + \underbrace{\lambda\|\omega\|^2}_{L_2 \text{ Regulizor}} \longrightarrow \min_{\omega}$$

Exam Tips:

2-1: When b and c are independent:

$$p(a|b) = \int p(a,c|b)\,dc$$

$$p(a,c|b) = \frac{p(a,c,b)}{p(b)} = \frac{p(a|b,c)\,p(b|c)\,p(c)}{p(b)}$$

if b and c are independent: $p(b|c) = p(b)$

$$\Rightarrow p(a|b) = \int p(a|b,c)\,p(c)\,dc$$

$\boxed{2\text{-}2}$

① $p(a,c|b) = \frac{p(a,b,c)}{p(b)}$, ② $p(c|a,b) = \frac{p(a,b,c)}{p(a,b)}$

$$\Rightarrow \frac{①}{②} = \frac{p(a,b)}{p(b)} = p(a|b)$$

Analytical Inference

$$\text{Posterior distribution: } P(\theta|X) = \frac{\overbrace{P(X|\theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(X)}_{\text{evidence}}}$$

- What is evidence $P(x)$?

*Maximum a posteriori principle:

- We try to find the value of parameters that maximizes the posterior probability:

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

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

avoid computing the evidence

* problems

- Not invariant to reparameterization
- Can't be used as a prior
- MAP is a solution to $L(\theta) = \mathbb{I}(\theta \neq \theta^*) \rightarrow \min_{\theta}$
- can't compute credible regions

Conjugate Distributions (another method to avoid computing evidence)

Fixed by model $\qquad\qquad\qquad\qquad$ our own choice

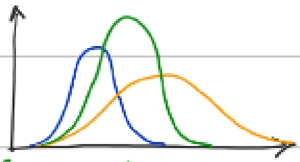
$$P(\theta|X) = \frac{P(X|\theta)\,P(\theta)}{P(X)}$$

Fixed by data

$-P(\theta)$ is conjugate to $P(X|\theta)$:

$$\mathcal{A}(v') \longrightarrow P(\theta|X) = \frac{P(X|\theta)\,\overbrace{P(\theta)}^{\mathcal{A}(v)}}{P(X)}$$

Example: Two Gaussians



$$\underbrace{P(\theta|X)}_{\mathcal{N}(\theta|a,b^2)} = \frac{\overbrace{P(X|\theta)}^{\mathcal{N}(X|\theta,\sigma^2)}\,\overbrace{P(\theta)}^{\mathcal{N}(\theta|m,s^2)}}{P(X)}$$

Example: Gamma distribution

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

$\gamma, a, b > 0$ $\quad \Gamma(n) = (n-1)!$

mean: $E(\gamma) = a/b$

$\text{Mode}[\gamma] = \frac{a-1}{b}$

$\text{Var}[\gamma] = a/b^2$

* Gamma distribution is conjugate to Normal distribution with respect to precision (inverse of variance $\gamma = \frac{1}{\sigma^2}$)

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

assume: $p(\gamma) \propto \gamma^{a-1} e^{-b\gamma} = \Gamma(\gamma|a,b)$

let's check: $p(\gamma|x) \propto p(x|\gamma)\, p(\gamma)$

$$p(\gamma|x) \propto \left(\gamma^{1/2} e^{-\frac{\gamma(x-\mu)^2}{2}}\right) \cdot \left(\gamma^{a-1} e^{-b\gamma}\right)$$

$$p(\gamma|x) \propto \gamma^{1/2+a-1} e^{-\gamma\left(b+\frac{(x-\mu)^2}{2}\right)}$$

$$\Rightarrow p(\gamma, x) = \Gamma\left(a+\tfrac{1}{2},\, b+\frac{(x-\mu)^2}{2}\right)$$

Example: Beta distribution (good for modeling final support)

$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}$$

$x \in [0,1]$, $a,b > 0$ $\qquad \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$

mean: $\mathbb{E}x = \frac{a}{a+b}$

$$\text{Mode}[x] = \frac{a-1}{a+b-2}$$

$$\text{Var}[x] = \frac{ab}{(a+b)^2(a+b-1)}$$

* Beta distribution is conjugate to Bernoulli likelihood

_ Bernoulli likelihood

$$p(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

and: $\quad p(\theta) = B(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$

$$p(\theta|X) \propto p(X|\theta)\, p(\theta)$$

$$p(\theta|X) \propto \theta^{N_1}(1-\theta)^{N_0} \cdot \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|X) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$: a Beta dist. func.

$\Rightarrow \quad p(\theta|X) = B(N_1+a, N_0+b)$

Summary

Pros:

— Exact posteriors

— Easy for online learning

Cons:

— Conjugate prior may be inadequate

Latent Variable Models

Pros:
- Simpler models (less edges)
- fewer parameters
- Latent variables are sometimes meaningful

Cons:
- Harder to work with

— Why using Bayesian method instead of standard regression approaches sometimes?

- There are some missing values
- To quantify uncertainty in predictions

— Latent Variable Models

- The connections of Bayesian model could be too complex.
- Therefore the table of probabilities could be very large
- The "Intelligence" is introduced to model/capture some features of each component.

* An Example:

$$p(x_1,\dots,x_5)=\sum_{I=1}^{100} p(x_1,\dots,x_5)|I)\,p(I)$$

Sum over the number of assigned intelligence

$$=\sum_{I=1}^{100} p(x_1|I)\dots p(x_5|I)\,p(I)$$

Five tables only are needed

Probabilistic Clustering

– Soft Clustering

$p(\text{cluster idx} \mid x)$ instead of cluster idx $= f(x)$

- It assumes that each point belongs to all clusters, but with different probabilities.

– Benefits of using...

- Handling missing values naturally
- Hyperparameter tuning
- Generating new data points

* Some Examples:

$\to p(x \mid \theta) = \mathcal{N}(x \mid \mu, \Sigma)$

$\theta = \{\mu, \Sigma\}$

– In case one Gaussian is not able to cover all data points one alternative is to use multiple distributions:

Gaussian Mixture Model (GMM)

$$p(x \mid \theta) = \pi_1 \mathcal{N}(x \mid \mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x \mid \mu_2, \Sigma_2) + \pi_3 \mathcal{N}(x \mid \mu_3, \Sigma_3)$$

$$\theta = \{\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3\}$$

How to train a GMM (estimation of params)?

- By finding maximum likelyhood: $\max_{\theta} p(X|\theta)$

$$\max_{\theta} p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

$$= \prod_{i=1}^{N} \left(\pi_1 \mathcal{N}(x_i|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \Sigma_2) + \dots\right)$$

N: number of data points

&

$$\pi_1 + \pi_2 + \pi_3 = 1 \; ; \; \pi_k \geqslant 0 \; ; \; k = 1, 2, 3$$

* For this optimization problem, tools like tensorflow can be used.

* The covariance matrix Σ is not arbitrary:
  It should be positive semi definite : $\Sigma_k \succ 0$

* Why not to use SGD in this case?
  * It is hard to follow some constraints
  * Expectation Maximization Algorithm, which can exploit the structure of the program, sometimes is much more faster and efficient.

Training GMM (Gaussian Mixture Model)

Q: How to do better than SGD?

$$p(x|\theta) = \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x|\mu_2, \Sigma_2) + \pi_3 \mathcal{N}(x|\mu_3, \Sigma_3) \qquad \text{(eq. 1)}$$

* Assign a latent variable "t" for each data point x.

$$t \longrightarrow x$$

* For this problem, $t = 1, 2, 3$ (three values), showing that which Gaussian this particular data point came from (which is unknown)
* it is reasonable to assume that "t" has prior distribution π.
* if we know that a data point comes from Gaussian number C:

$$p(t=c|\theta) = \pi_c \longleftarrow \text{prior}$$

* The density of data point, given the cluster ~~x~~:

$$p(x|t=c, \theta) = \mathcal{N}(x|\mu_c, \Sigma_c)$$

* Rule of sum for probabilities:

$$p(x|\theta) = \sum_{c=1}^{3} p(x|t=c, \theta)\, p(t=c|\theta) \qquad \text{(eq. 2)}$$

Marginalizing "t"

* eq. 1 is exactly the same as eq. 2.

Expectation Maximization

* How to train this latent variable model?

- How to estimate θ?

- From training data set, if "t" is known (for example):

$$p(x \mid t=1, \theta) = \mathcal{N}(x \mid \mu_1, \sigma_1^2)$$

$$\mu_1 = \frac{\sum_{blue} x_i}{\# \text{ of blue points}}, \quad \sigma_1^2 = \frac{\sum_{blue} (x_i - \mu_1)^2}{\# \text{ of blue points}}$$

$$\mu_1 = \frac{\sum_i p(t_i=1 \mid x_i, \theta)\, x_i}{\sum_i p(t_i=1 \mid x_i, \theta)}$$

$$\sigma_1^2 = \frac{\sum_i p(t_i=1 \mid x_i, \theta)\, (x_i - \mu_1)^2}{\sum_i p(t_i=1 \mid x_i, \theta)}$$

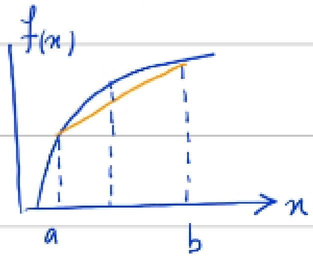
can also be calculated from posterior

* How can we choose the best run among several training attempts with different random initialization?

- choose the one with the highest training log-likelihood
- choose the one with the highest validation log-like...

General Form of Expectation Maximization

* Concave Functions



for any a, b, α : $f(\alpha a + (1-\alpha)b) \geqslant \alpha f(a) + (1-\alpha) f(b)$

$0 \leqslant \alpha \leqslant 1$

* Jensen's inequality

If "f" is a concave function and

$\alpha_1 + \alpha_2 + \alpha_3 = 1$; $\alpha_k \geqslant 0$

$$\Rightarrow f(\underbrace{\alpha_1 a_1 + \alpha_2 a_2 + \alpha_3 a_3}_{\mathbb{E}_{p(t)} t}) \geqslant \underbrace{\alpha_1 f(a_1) + \alpha_2 f(a_2) + \alpha_3 f(a_3)}_{\mathbb{E}_{p(t)} f(t)}$$

$p(t=a_1) = \alpha_1$

$p(t=a_2) = \alpha_2$

$p(t=a_3) = \alpha_3$

$$\Rightarrow \quad f(\mathbb{E}_{p(t)} t) \geqslant \mathbb{E}_{p(t)} f(t)$$

Kullback-Leibler Divergence

↳ It's a way to measure difference between two probabilistic distributions.

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

$\int$ is Expected value of $q(x)$ — a measure of diff

* $KL(q||p) \neq KL(p||q)$
* $KL(q||q) = 0$
* $KL(q||p) \geqslant 0$

proof: $-KL(q||p) = \mathbb{E}_q(-\log q/p) = \mathbb{E}(\log p/q)$

$$\leqslant \log(\mathbb{E}_q\, p/q) = \log \int q(x) \frac{p(x)}{q(x)} = 0$$

General form of Expectation Maximization

$*$ to find out $\max_\theta p(X|\theta)$, max of likelihood

\# of data points

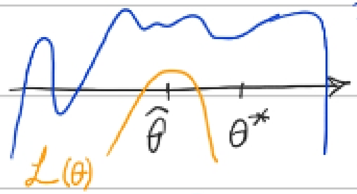
$$\max_\theta \log p(X|\theta) = \log \prod_{i=1}^{N} p(x_i|\theta)$$

$$\log p(X|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta)$$

$$\text{(eq.1)} \quad = \sum_{i=1}^{N} \log \sum_{c=1}^{3} p(x_i, t_i = c|\theta) \geq \mathcal{L}(\theta)$$

a lower bound - Jensen's inequality

e.g.



$*$ one lower bound is not enough, so we define a family of lower bounds.

$$\text{(eq.1)} = \sum_{i=1}^{N} \log \sum_{c=1}^{3} \frac{q(t_i=c)}{q(t_i=c)} p(x_i, t_i = c|\theta)$$

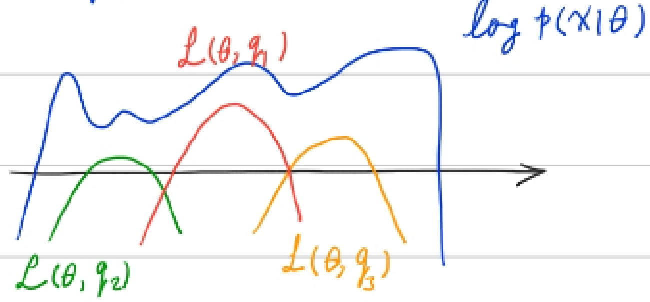
$$\boxed{\log\left(\sum_c \alpha_c v_c\right) \geq \sum_c \alpha_c \log(v_c)}$$

Jensen's inequality

$\alpha_c : q(t_i = c)$ $\qquad v_c : \dfrac{p(x_i, t_i=c|\theta)}{q(t_i=c)}$

$$\Rightarrow \text{(eq.1)} \geq \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i=c) \log \frac{p(x_i, t_i=c|\theta)}{q(t_i=c)}$$

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



* Iteration Steps:

– For θ_k, find $\mathcal{L}(\theta, q)$ which is maximum at this current point θ_k. / find q such that the value of a lower bound of the point θ_k and q is maximum.
(check PDF of slides)

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

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

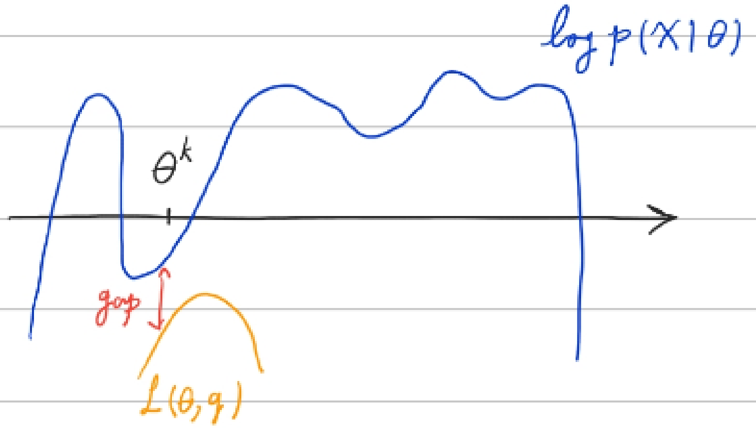
E-Step Details

$$\log p(X|\theta) \geqslant \mathcal{L}(\theta, q)$$

q: itself is a distribution

E-Step: $\max\limits_{q} \mathcal{L}(\theta^k, q)$

1) choose the one which has the higest value at θ_k.



gap:

$$\log p(X|\theta) - \mathcal{L}(\theta, q) = \sum_i KL\left(q(t_i) \,||\, p(t_i | x_i, \theta)\right)$$

proof:

- we assume that the data-set consists of the objects that are independent of given parameters.

$$= \sum_{i=1}^{N} \log p(x_i|\theta) - \sum_{i=1}^{N} \log \sum_{c=1}^{3} q(t_i=c) \frac{p(x_i, t_i=c|\theta)}{q(t_i=c)}$$

$$= \sum_{i=1}^{N} \left(\log p(x_i|\theta) \cdot \underbrace{\sum_{c} q(t_i=c)}_{1} - \sum_{c=1}^{3} q(t_i=c) \cdot \log \ldots \right)$$

$$= \sum_{i=1}^{N}\sum_{c}^{3} q(t_i=c)\left(\log p(x_i|\theta) - \log \frac{p(x_i, t_i=c|\theta)}{q(t_i=c)}\right)$$

$$= \sum_{i=1}^{N}\sum_{c}^{3} q(t_i=c) \log \frac{p(x_i|\theta)\, q(t_i=c)}{\underset{\substack{\downarrow \\ p(t_i=c|x_i,\theta)\, p(x_i|\theta)}}{p(x_i, t_i=c|\theta)}}$$

$$= \sum_{i=1}^{N}\underbrace{\sum_{c}^{3} q(t_i=c) \log \frac{q(t_i=c)}{p(t_i=c|x_i,\theta)}}_{KL\left(q(t_i)\,\|\, p(t_i|x_i,\theta)\right)}$$

$$= \sum_{i=1}^{N} KL\left(q(t_i)\,\|\, p(t_i|x_i,\theta)\right) \qquad (Eq.\ 2)$$

$$GAP = \log p(X|\theta) - \mathcal{L}(\theta, q)$$

max q   min q   min q

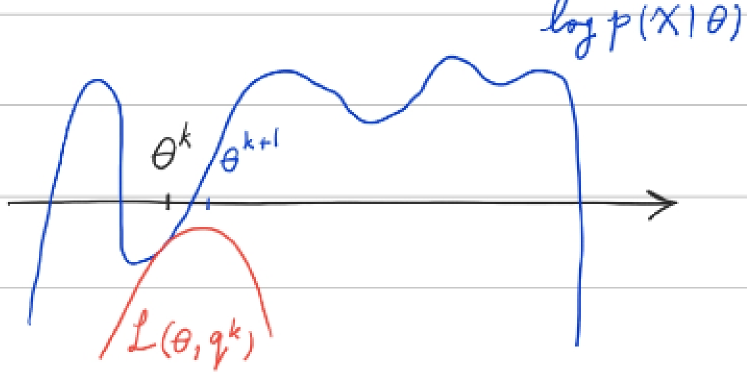
$$\Rightarrow \min_{q}\ (Eq.\ 2) \geq 0 \Longrightarrow q(t_i) = p(t_i|x_i,\theta)$$

for global optimal

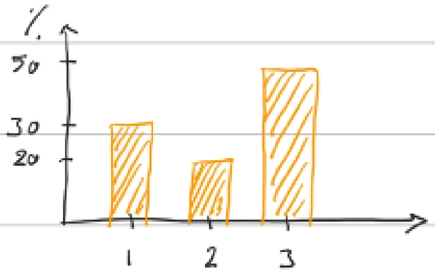
$\ast$ E-Step

$$\underset{q(t_i)}{\arg\max}\ \mathcal{L}(\theta^k, q) = p(t_i|x_i,\theta)$$

$\Longrightarrow$



Example: EM for Discrete Mixture, E-Step



$*$ Assumtion:

- probability of x_i: mixture of two distributions

$$p(x_i) = \gamma p_1(x_i) + (1-\gamma) p_2(x_i)$$

- lets say that:

	1	2	3
p_1	α	$1-\alpha$	0
p_2	0	$1-\beta$	β

- objective: to estimate α, β, and γ via EM method.

1) parameters initialization:

$$\alpha_0 = \beta_0 = \gamma_0 = 0.5$$

2) Define the latent variable t_i for each x_i

$$x_i \leftarrow t_i$$

- t_i can just take two values

e.g.: $p(t_i=1) = \gamma$, $p(t_i=2) = 1-\gamma, \dots$

$\Rightarrow p(x_i | t_i=2) = p_2(x_i)$

- E-Step:

Finding out the posterior distribution on the latent variable t_i:

$\Rightarrow \quad q(t_i=c) = p(t_i=c \mid x_i)$

Start with:

$$p(t_i=1 \mid x_i=1) = \frac{p(x_i=1 \mid t_i=1)\, p(t_i=1)}{p(x_i=1 \mid t_i=1)\, p(t_i=1) + p(x_i=1 \mid t_i=2)\, p(t_i=2)}$$

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

$$p(t_i=1 \mid x_i=3) = \frac{p(x_i=3 \mid t_i=1)\, p(t_i=1)}{p(x_i=3 \mid t_i=1)\, p(t_i=1) + p(x_i=3 \mid t_i=2)\, p(t_i=2)}$$

$$= 0$$

$$p(t_i=1 \mid x_i=2) = \ldots = \frac{0.5 \times 0.5}{0.5 \times 0.5 + 0.5 \times 0.5} = 0.5$$

Example: EM for discrete mixture: M-Step

Objective:

$$\max_{\alpha,\beta,\gamma} \sum_{i=1}^{N} \mathbb{E}_{q(t_i)} \log \overbrace{p(x_i|t_i)\,p(t_i)}^{p(x_i,t_i)}$$

Summary of E-Step:

$$q(t_i=1) = p(t_i=1|x_i) = \begin{cases} 1 & ; x_i=1 \\ 0.5 & ; x_i=2 \\ 0 & ; x_i=3 \end{cases}$$

$$q(t_i=2) = 1 - q(t_i=1)$$

$$= \sum_{i=1}^{N} q(t_i=1) \log p(x_i|t_i=1)\,\gamma + \sum_{i=1}^{N} q(t_i=2) \log p_2(x_i)(1-\gamma)$$

$*$ assumption: $N_1=30$, $N_2=20$, $N_3=60$

$$= 30 \cdot \underbrace{p(t_i=1|x_i=1)}_{1} \log \alpha\gamma + 20\cdot 0.5 \cdot \log(1-\alpha)\,\gamma$$

$$+ \cancelto{0}{60\cdot 0 \cdot \log 0} + \cancelto{0}{30\, p(t_i=2|x_i=1)} \cdots + 20\cdot 0.5 \cdot \log(1-\beta)(1-\gamma)$$

$$+ 60\cdot 1 \cdot \log \beta(1-\gamma)$$

maximizing wrt α: $\longrightarrow$ just consider terms depending on α

$$f_1(\alpha) = 30 \log \alpha + 10 \log(1-\alpha) + \text{const}(\alpha) \implies \max f_1(\alpha)$$

$$\frac{\partial f_1(\alpha)}{\partial \alpha} = 0 \implies 30 \frac{1}{\alpha} + 10 \frac{(-1)}{1-\alpha} = 0 \implies \dots$$

summary: $\alpha = \frac{3}{4}$, $\beta = \frac{6}{7}$, $\gamma = \frac{4}{11}$

meaning:

	1	2	3
P_1	α	$1-\alpha$	0
P_2	0	$1-\beta$	β

General EM for GMM

Objective: How to apply the EM algorithm to some concrete latent variable models.