Bayesian Methods for ML

X & Y are ⊥:

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

(Joint, Marginal, Marginal)

Conditional Prob

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

(Conditional; Joint; Marginal)

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

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

OR $P(\text{1st Grp} \mid \text{2nd Grp}) = P(\text{2nd Grp} \mid \text{1st Grp})$

$$= \frac{P(\text{1st Grp})}{P(\text{2nd Grp})}$$

Chain Rule

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

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

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

Sum Rule

$$P(X) = \int_{-\infty}^{\infty} P(X,Y)\,dY$$

↑ Marginalization

Bayes theorem

Θ – parameters
X – observations

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

- $P(\Theta|X)$: Posterior (probability of params after we observe the data)
- $P(X,\Theta)$: Joint (Chain Rule)
- $P(X)$: Evidence
- $P(X|\Theta)$: likelihood
- $P(\Theta)$: Prior (prior knowledge about params, for instance some params are distributed around zero)

Bayesian Vs Frequentist

Frequentist

Bayesian (See W1-L2)

Classification

Training:

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

← Posterior

Prediction: (Marginalize "θ")

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

In Bayesian Approach, prediction is a weighted average of output of our model for all possible values of parameters.

* Bayes formula can also lead to regularisation

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

$P(\theta)$ = prior & can act as regularizer

* Bayesian methods are good for online learning

For point x_K,

$$P(\theta | x_K) = \frac{P(x_K | \theta)\, P(\theta)_{1<-1}}{P(x_K)}$$

↑ Posterior ← Likelihood ← Prior

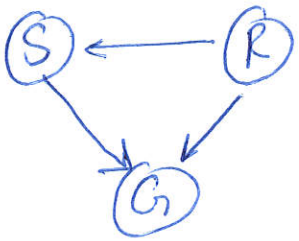
After computing posterior $P(\theta|x_K)$; lets say we receive new point x_{K+1}

for x_{K+1} , $P(\theta|x_{K+1}) = \frac{P(x_{K+1}|\theta)\, P_K(\theta)}{P(x_{K+1})}$

$P_K(\theta)$ can be taken as $P(\theta|x_K)$ i.e priors can be updated when new point arrives.

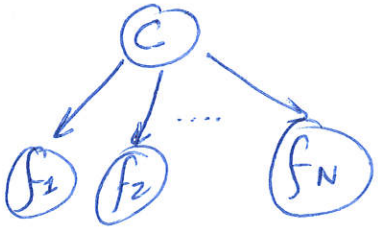
How to define Model (23_101)

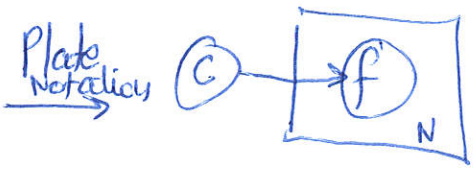
Model: Joint Probability over all variables



$$P(S,R,G_1) = P(G_1|S,R)\, P(S|R)\, P(R)$$

Naive Bayes Classifier



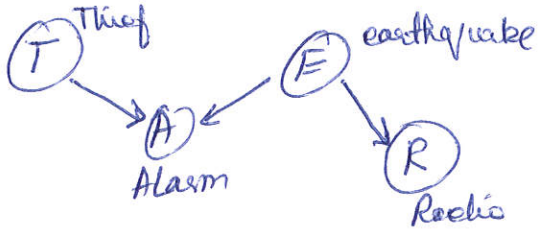


Joint Prob:

$$P(C, f_1, f_2 \cdots f_N) = P(C) \prod_{i=1}^{N} P(f_i|C)$$

* Bayesian Networks can not have/contain directed cycles. We can't have interdependent variables. For those cases we can ~~join~~ either joint random variables into one random variable or use MRF.

101-23: Thief and Alarm



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

To fully define our model, we need to define these 4-probabilities
i.e $P(t)\ P(e)\ P(a|t,e)\ P(r|e)$

$P(t)$ & $P(e)$

Priors	
$P(T=1)$	$1/1000$
$P(E=1)$	$1/100$

$P(a|t,e)$

$P(A=1\|T,E)$	$E=0$	$E=1$
No Thief: $T=0$	0	0.1
Thief: $T=1$	1	1

$P(r|e)$ → radio report given earthquake

$P(r\|e)$	
$E=0$	0
$E=1$	0.5

Calculate the following

*) $P(T|A)$ = thief given alarm

Using Bayes formula

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

Use sum rule

$$= \frac{P(T,A,E) + P(T,A,\bar{E})}{P(T,A,E)+P(T,A,\bar{E})+P(\bar{T},A,\bar{E})+P(\bar{T},A,E)} \rightarrow \text{eq(A)}$$

In eq(A):

i) $P(T,A,E) = P(A|T,E)\,P(T)\,P(E)$
$= (1)\ (10^{-3})(10^{-2})$
$\Rightarrow P(T,A,E) = 10^{-5}$

* $P(T,A,\bar{E}) = P(A|T,\bar{E})\,P(T)\,P(\bar{E})$
$= (1)(10^{-3})(1-0.01)$
$= 10^{-3}(0.99)$
$= 0.00099$

ii) $P(\bar{T},A,\bar{E}) = P(A|\bar{T},\bar{E})\,P(\bar{T})\,P(\bar{E})$
$= (0)\ (\ )\ (\ )$
$= 0$

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

($P(T,A,E) = 10^{-5}$; $P(\bar{T},A,\bar{E}) \rightarrow 0$) $\approx 50\%$

$\not$ $P(T|A,R)$ = ratio of joint Probabilities

$$= \frac{P(A,T,R)}{P(A,R)} \quad \approx 1\%$$

Add missing variables and using sum rule

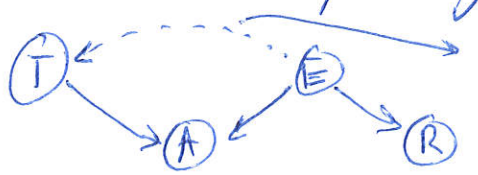
$$= \frac{P(\not{A}T,A,R,E) + P(\bar{T},A,R,\bar{E})}{P(A,R,T,E) + P(A,R,\bar{T},\bar{E}) + P(A,R,\bar{T},E) + P(A,R,\bar{T},\bar{E})} \rightarrow \text{Eq(B)}$$

$P(\not{A}T,A,R,E) \rightarrow 0$; $P(A,R,\bar{T},\bar{E}) \rightarrow 0$

In Eq(B) :

i) $P(T,A,R,E) = P(t)\,P(a|t,e)\,P(r|e)\,P(e)$

1% for $P(\text{thief}|\text{Alarm, radio report})$ suggests that our graphical model should contain some dependency, like:



i.e there are more thiefs when earthquake happens

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

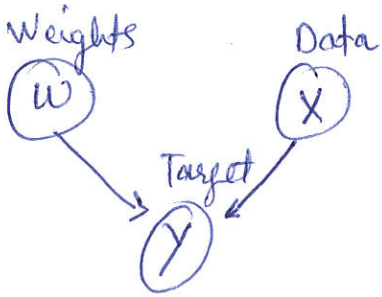
Least squares problem

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

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

Linear Regression in light of Bayesian N/w's

MODEL



Three Random Variables
i) w
ii) X
iii) Y

We are not interested in modelling the data (X), therefore we can write the joint probability of weight & target given data as follows:

$$P(w, y|X) = P(y|X, w)P(w) \quad [\text{See the model above}]$$

Note that $P(w, y, X) = P(w, y|X)$ since we are not modelling data (X)

In such case: $P(w, y, X) = P(y|w, X)\,P(w)\,P(X)$

Since we don't model X,

$$\Rightarrow P(w, y, X) = P(w, y|X) = P(y|w, X)\,P(w)$$

Now, we need to define two distributions: $P(y|w, X)$ & $P(w)$. Let them be Normal distribution

So $P(y|w, X) = \mathcal{N}(y|\, w^T X, \sigma^2 I)$ — mean, variance

The probability of target given weights & observations is gaussian centred at prediction $w^T X$ & variance $\sigma^2 I$

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

Now, lets train linear regression given the following formulas:

$$P(w, y|X) = P(y|X, w)P(w) \longrightarrow \text{Eq(A)}$$

$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 I) \longrightarrow \text{Eq(B)}$$

$$P(w) = \mathcal{N}(w|0, \gamma^2 I) \longrightarrow \text{Eq(C)}$$

Lets calculate the posterior probability of $w/y, X$

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

Since we are not modeling X

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

$$P(w|y,x) = \frac{P(w,y,x)}{P(y,x)} = \frac{P(w,y|X)}{P(y|X)} = \dots$$

we want to find $\max_w P(w|y, X)$

$$\Rightarrow w^* = \max_w \frac{P(y, w|X)}{P(y|X)}$$

* when maximizing $\frac{P(y,w|X)}{P(y|X)}$, $P(y,w|X)$ depends on w.

In order to $\max_w P(w/y, X)$, we can maximize $P(y, w/X)$.

$$\Rightarrow w^* = \max_w P(y, w|X)$$

$$= \max_w P(y|X, w)p(w) \longrightarrow \text{From eq(A)}$$

Applying log

$$w^* = \max_w \log\left(P(y|X, w)\, p(w)\right)$$

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

$$= \max_w \log C_1 \cdot \exp\left(-\frac{1}{2}(y - w^T X)^T [\sigma^2 I]^{-1} (y - w^T X)\right) + \log C_2 \cdot \exp\left(\frac{-1}{2} w^T [\gamma^2 I]^{-1} w\right)$$

$$= \max_w \left[-\frac{1}{2\sigma^2}(y - w^T X)^T (y - w^T X) - \frac{1}{2\gamma^2} w^T w\right]$$

$$= \max_w \left[-\frac{1}{2\sigma^2}\|y - w^T X\|^2 - \frac{1}{2\gamma^2}\|w\|^2\right]$$

Multiply by $-1 \times 2\sigma^2$

$$= \min_w \; \underbrace{\|y - w^T x\|^2}_{\text{Sum of squares}} + \lambda \underbrace{\|w\|^2}_{\text{L2-regularizer}}$$

$$\Rightarrow w^* = \max_w p(w|y,X) = \min_w \left[\|y - w^T x\|^2 + \lambda \|w\|^2 \right]$$

So, adding a normal prior on the weights $[p(w)]$, we turned from the least squares problem to the L2-regularized least-squares problem. i.e solving a least-squared problem with L2-regularizer is equivalent of finding the MAP estimate for w with prior distribution $\mathcal{N}(w|0, \gamma I)$.

Conjugate Priors ~~(W1-L1)~~

Analytical Inference (W1-L1)

* Bayesian Inference has some complications. One of them is $P(X)$ i.e computation of Evidence.

Posterior Distribution

$$P(\theta|X) = \frac{\overbrace{P(X|\theta)}^{\text{Likelihood}} \overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{P(X)}_{\text{evidence}}}$$

Computing this evidence is hard & complicated. For instance, in images if we model $P(X)$ then we can draw new images. Generally computing $P(X)$ is hard. We will see how can we avoid calculation of evidence during inference. It is called Maximum a posteriori principle.

Maximum A Posteriori

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

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

$$\theta_{MP} = \arg\max_\theta P(X|\theta)P(\theta)$$: This is independent of "X" so we can avoid computing $P(X)$

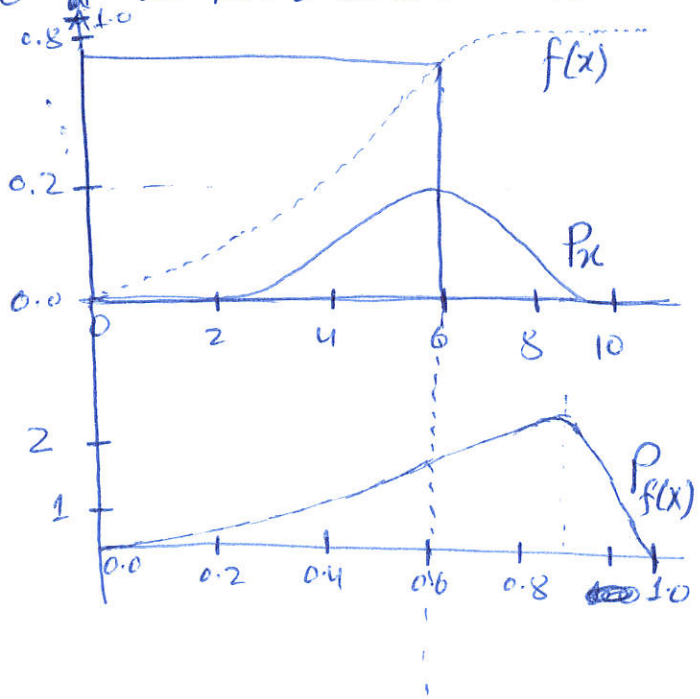
Thus it becomes an optimization problem.

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

MAP PROBLEMS

Maximum a posteriori has its own problem. It is not invariant to reparametrization.

i) Imagine we have a random variable "x" having Gaussian distribution as follows. We will apply sigmoid to it $[f(x)]$ and get $P_{f(x)}$ as an output. Note that the position of "MAXIMUM" changes ~~therefore we can't use it as Prior distribution~~.



ii) Next problem is we can't use it as a prior

$$P_K(\theta) = \frac{P(x_K|\theta)\ P_{K-1}(\theta)}{P(x_K)}$$

If we try to use MAP as prior, we will get the delta function and so we will not get any new information

$$P_K(\theta) = \frac{P(x_K|\theta)\ \delta(\theta-\theta_{MP})}{P(x_K)} = \delta(\theta-\theta_{MP})$$

Conjugate Distributions

Bayes formula:

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

$P(X|\theta)$ → Fixed by model

$P(\theta)$ → One own choice

$P(X)$ → Fixed by data

Lets select $P(\theta)$ in a way that it becomes easier to compute the posterior:

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

$P(\theta|X)$ → $A(v')$; $P(\theta)$ → $A(v)$

The prior is conjugate to the likelihood if the prior & the posterior lie in the same family of distributions. For instance if the prior is normal with some "μ" & "σ" then we would expect the posterior to also be normal but with some other mean & variance.

Example

$P(X|\theta) = \mathcal{N}(X|\theta, \sigma^2)$

$A(v) = ??$

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

$P(X|\theta)$ → $\mathcal{N}(X|\theta,\sigma^2)$; $P(\theta)$ → $A(v)$; $P(\theta|X)$ → $A(v')$

what should be the conjugate prior $A(v)$?

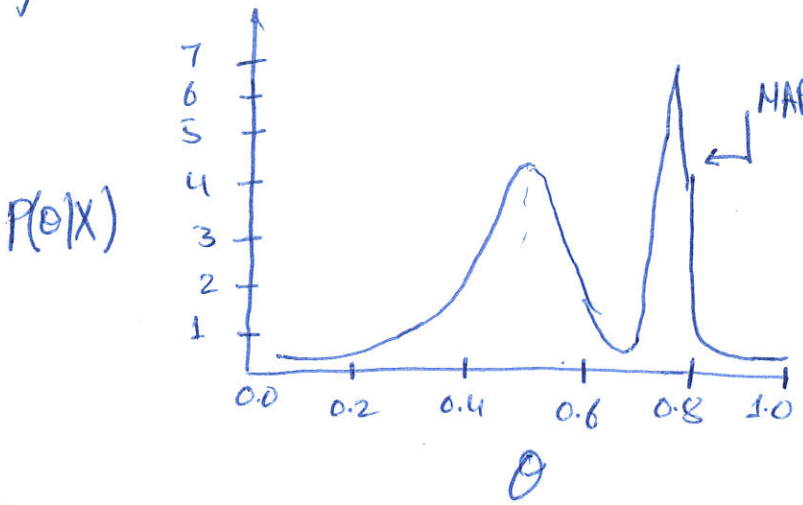
A: If we select $A(v)$ as a normal distribution then $P(\theta|X)$ will also have normal distribution. It is because product of two normal distributions $P(X|\theta)$ & $P(\theta)$ yields normal distribution

i.e

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

$P(X|\theta)$ → $\mathcal{N}(X|\theta,\sigma^2)$; $P(\theta)$ → $\mathcal{N}(\theta|m,s^2)$; $P(\theta|X)$ → $\mathcal{N}(\theta|a,b^2)$

iii) Another problem is that Maximum Aposteriori estimation is actually an untypical point since there may be not enough probability density around it.



* MAP is a solution to $L(\theta) = \Pi[\theta \neq \theta^*] \to \min_{\theta}$

It also equals to the result of the minimization of the indicator that you do not end up in the true ideal parameter value.

We could use other functions for e.g the squared/Absolute error. Those would lead to the mean or median of the posterior distribution

Objectives	Solution
$L(\theta) = \Pi[\theta \neq \theta^*] \to \min_{\theta}$	Mode
$L(\theta) = E(\theta - \theta^*)^2 \to \min_{\theta}$	Mean
$L(\theta) = E\lvert\theta - \theta^*\rvert \to \min_{\theta}$	Median

And if you choose these functions, we need to calculate evidence. That's what we are trying to avoid.

iv) finally we can't compute credible regions i.e if $\theta_{MP} = 12.53$ we can't say how confident are we.

Summarizing Maximum Aposteriori Estimation

PROS

* EASY to Compute

CONS

* Not invariant to reparametrization
* Can't use as a prior
* Find's untypical point
* Can't compute credible regions.

* Another way to avoid computing evidence : conjugate distributions

Example

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)} = \frac{\mathcal{N}(x|\theta,1)\,\mathcal{N}(\theta|0,1)}{p(x)}$$

$$p(\theta|x) \propto e^{-\frac{1}{2}(x-\theta)^2} e^{-\frac{1}{2}\theta^2}$$

$$p(\theta|x) \propto e^{-(\theta-\frac{x}{2})^2}$$

i.e $p(\theta|x)$ is Normal with mean $\frac{x}{2}$ & variance $\frac{1}{2}$. This also implies prior is conjugate to likelihood.

Question $p(\theta)$ is Gaussian, $P(x|\theta)$ is also Gaussian with mean 0 & variance σ^2. Here $P(\theta)$ is not conjugate to $P(x|\theta)$.

W1-P2-23 Example: Normal, precision

- More examples for conjugate priors
- Gamma Distribution (parametrized by a, b)

Its PDF is:

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

$y, a, b > 0$ (Gamma distribution is the distribution over +ve axis)

$\Gamma(n) = (n-1)!$

Statistics

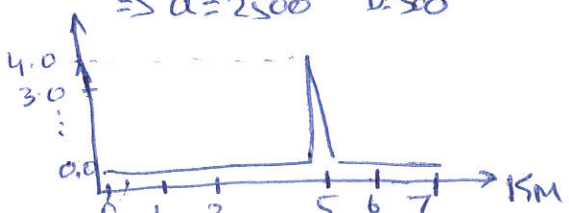
$$\underbrace{E[y] = a/b}_{\text{mean}} \quad \text{Mode}[y] = \frac{a-1}{b} \quad \text{Var}[y] = a/b^2$$

Support of Gamma ~~function~~ distribution $= [0, \infty)$

Example You run 5Km ± 100 m a day. (Model it by $\Gamma(y|a,b)$)

$E[x] = a/b = 5 \quad \text{Var}[x] = a/b^2 = 0.1^2$

$\Rightarrow a = 2500 \quad b = 50$



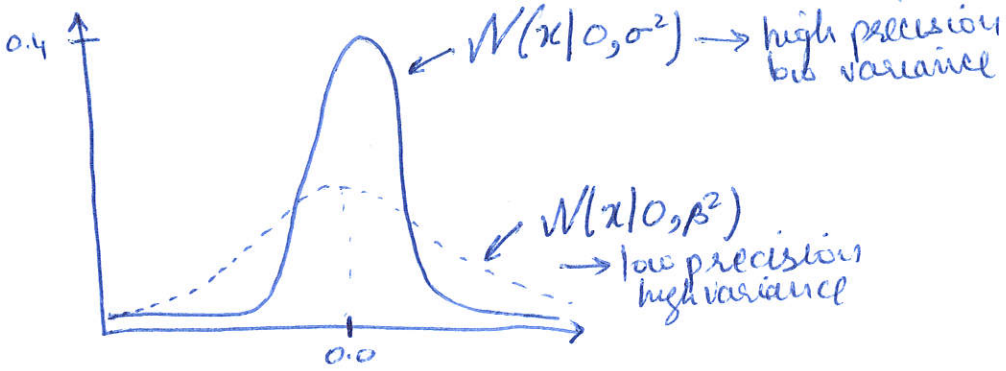
W1-P2-L4 Example: Normal, precision

Statement: The gamma distribution is conjugate to the normal w.r.t precision.

$$\text{Precision } (\gamma) = \frac{1}{\text{variance } (\sigma^2)}$$

OR

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



Precision

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

As $\sigma^2 = 1/\gamma$

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

Now, what is the conjugate prior w.r.t to the precision?

Ans: Here is functional form

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

Dropping all constants

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

$$\propto \gamma^{1/2} e^{-b\gamma}$$

Example: Bernoulli

For final example, we will take Beta distribution

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

$x \in [0,1]$

$a, b > 0$

You can get multiple distributions like constant, U-shaped, uniform.

Statistics

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

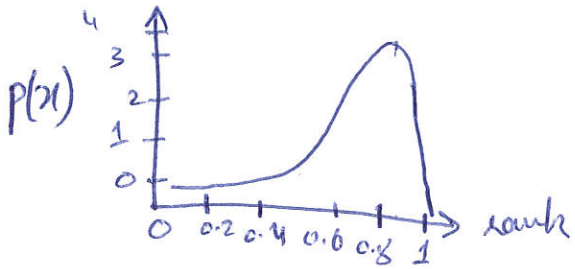
Support of beta distribution = $[0,1]$

Now: Use Beta Distribution to model our favourite movie ranks

Movie rank is 0.8 ± 0.1

$$Ex = \frac{a}{a+b} = 0.8$$

$$Var[x] = 0.1^2 \Rightarrow a = 12, b = 3$$



Actually, the Beta distribution is conjugate of the Bernoulli likelihood

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

$N_1 \rightarrow$ # of ones in dataset
$N_0 \rightarrow$ # of zeros

Lets select Beta distribution as prior

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

Thus posterior:

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

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

$$\Rightarrow p(\theta|x) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

This is Beta distribution with following params

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

Lets try finding conjugate distribution in the following way:

$$p(y) \propto y^{1/2} e^{-by}$$

Use the Bayes formula:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto \underbrace{y\, e^{-y\left(b+\frac{(x-\mu)^2}{2}\right)}}_{\text{After dropping constants \& rearranging terms}}$$

As we have seen that y has power "1" instead of $(1/2) \rightarrow$ this means it does not lie in the same family of distribution

Now, what if we choose gamma distribution. We will have another parametre that would allow us to vary the power of gamma

$$\text{As } p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$\text{where } \mathcal{N}(x|\mu, y^{-1}) = \frac{\sqrt{y}}{\sqrt{2\pi}} e^{-y\frac{(x-\mu)^2}{2}} \propto y^{1/2} e^{-by}$$

$$p(y) \propto y^{a-1} e^{-by}$$

This $p(y)$ is gamma distribution with following params:

$$p(y) = \Gamma(y|a,b)$$

Now, lets try to compute our posterior $p(y|x)$:

our prior is $p(y)$: $p(y) = \Gamma(y|a,b) \propto y^{a-1} e^{-by}$

our posterior $p(y|x) \propto p(x|y)p(y)$ [proportional to likelihood & prior]

Dropping all constants yields:

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

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

$$p(y|x) = \Gamma\left(\underbrace{a+\frac{1}{2}}_{\text{mean}}, \underbrace{b+\frac{(x-\mu)^2}{2}}_{\text{variance}}\right)$$

So we avoid computing the evidence by choosing the conjugate prior and calculate posterior $p(y|x)$ successfully.

Overall Picture

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

We would like to compute posterior $P(\theta|X)$ but computing evidence $P(X)$ makes it hard. Thus we choose conjugate Priors that helped us computing $P(\theta|X)$ without computing the evidence $P(X)$.

Conjugate Pros & Cons

PROS

* Exact posterior is computed using conjugate
* It is easy for online-learning
  Eg $P(\theta|X) = B(N_1+a, N_0+b)$

CONS

* Conjugate Prior may be inadequate for some models. Therefore in the next week, we will see more advanced techniques to compute the full posterior or sometimes the approximate posterior.