

Bayesian Methods for ML

①

$X \perp \!\!\! \perp Y$ are \perp :

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

Joint Marginal Marginal

Conditional Prob

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

conditional Joint Marginal

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

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

Chain Rule

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

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

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

$$\text{OR } P(\text{1st Grep} | \text{and Grep}) = \frac{P(\text{and Grep} | \text{1st Grep})}{P(\text{1st Grep})}$$

$$= \frac{P(\text{1st Grep})}{P(\text{and Grep})}$$

Sum Rule

$$P(X) = \int_{-\infty}^{\infty} p(x, y) dy$$

↑
Marginalization

Bayes theorem

θ - parameters

X - observations

$$P(\theta|X) = \frac{P(X|\theta) \underbrace{P(\theta)}_{\text{Prior}}}{P(X) \underbrace{\int P(X|\theta)P(\theta)d\theta}_{\text{Evidence}}} = \frac{\underbrace{P(X|\theta)P(\theta)}_{\text{Likelihood}}}{P(X)}$$

↑
Posterior

↑
chain Rule
Joint

↑
Evidence

↑
Likelihood

↑
Prior (prior knowledge about params)
for instance some means are distributed around zero

(probability of params after we observe the data)



Bayesian Vs frequentist

(3)

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
X_{tr}, y_{tr}
↓
θ

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 regularization

$$P(\theta | X) = \frac{P(X | \theta) P(\theta)}{P(X)} \quad P(\theta) = \text{prior \& can act as regularizer}$$

* Bayesian methods are good for online learning

$$\text{For point } x_K, \quad P(\theta | x_K) = \frac{P(x_K | \theta) P(\theta) \leftarrow \text{likelihood}}{\underset{\text{Posterior}}{\underbrace{P(x_K)}} \leftarrow \text{Prior}}$$

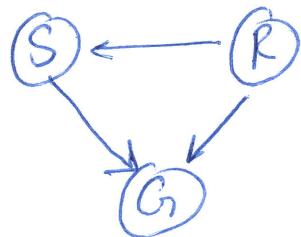
After computing posterior $P(\theta|x_R)$; lets say we receive new point x_{R+1}

$$\text{for } x_{R+1}, P(\theta|x_{R+1}) = \frac{P(x_{R+1}|\theta) P_\theta(\theta)}{P(x_{R+1})}$$

$P_\theta(\theta)$ can be taken as $P(\theta|x_R)$ i.e. prior can be updated when new point arrives.

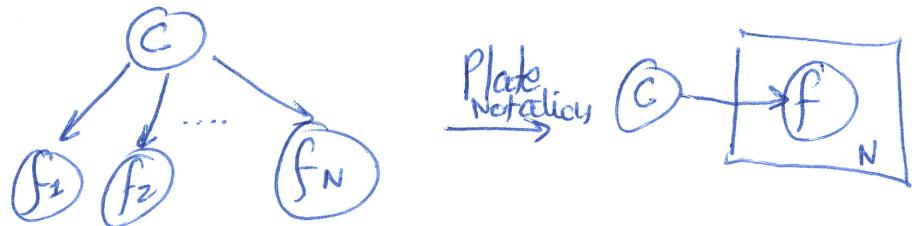
How to define Model (L3-10)

Model: Joint Probability over all variables



$$P(S, R, G_1, G_2) = P(G_1|S, R) P(S|R) P(R)$$

Naive Bayes Classifier



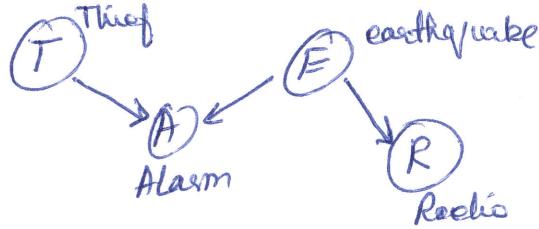
Joint Prob:

$$P(C, f_1, f_2, \dots, f_n) = P(C) \prod_{i=1}^n P(f_i|C)$$

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

W1-L3: Thief and Alarm

(5)



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

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

Prior	
	$P(t) \& P(e)$
$P(T=1)$	$\frac{1}{1000}$
$P(E=1)$	$\frac{1}{100}$

		$P(a t, e)$	
		$E=0$	$E=1$
No Thief: $T=0$		0	0.1
Thief: $T=1$		1	1

$P(r e) \rightarrow$ radio report given earthq/wake	
$P(r e)$	
$E=0$	0
$E=1$	0.5

Calculate the following

i) $P(T|A) = \text{thief given alarm}$

Using Bayes formula

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

Use sum rule

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

In eq(A):

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

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

$$\begin{aligned} ii) \quad P(\bar{T}, A, \bar{E}) &= P(A|\bar{T}, \bar{E})P(\bar{T})P(\bar{E}) \\ &= (0) () () \\ &= 0 \end{aligned}$$

$$\Rightarrow P(T|A) = \frac{P(T, A, E) + P(T, A, \bar{E})}{P(T, A, E) + P(T, A, \bar{E}) + P(\bar{T}, A, E) + P(\bar{T}, A, \bar{E})} \rightarrow 0 \approx 50\%$$

$$* P(T|A, R) = \text{ratio of joint Probabilities}$$

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

≈ 1%

Add missing variables and using sum rule

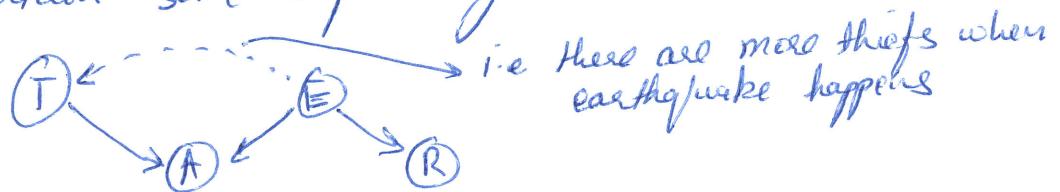
$$= \frac{P(\bar{A}T, A, R, E) + P(T, A, R, \bar{E})}{P(A, R, \bar{T}, E) + P(A, R, T, \bar{E}) + P(A, \bar{R}, \bar{T}, E) + P(A, \bar{R}, T, \bar{E})} \rightarrow \text{Eq(B)}$$

○ ○ ○ ○

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:

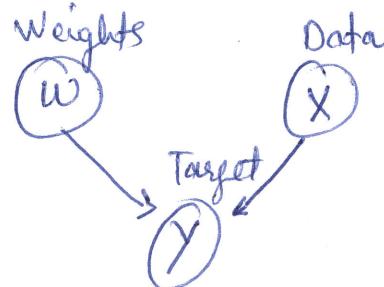


$$P(\bar{T}, A, R, E) = P(A|\bar{T}, E) P(R|E) P(T|E) P(E)$$

Least squares problem

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

$$\hat{w} = \operatorname{argmin}_w L(w)$$

Linear Regression in light of Bayesian N/WgMODEL

Three Random Variables

$$\begin{array}{l} \stackrel{(1)}{\sim} w \\ \stackrel{(2)}{\sim} X \\ \stackrel{(3)}{\sim} y \end{array}$$

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

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

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 $P(y | w, X)$. Let them be Normal distribution

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

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

The probability of
target given weight
& observation is
gaussian centred at
prediction $w^T X$ &
variance σ^2

Now, let's train linear regression given the following formulas:

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

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

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

Let's 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

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

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

$$\Rightarrow w^* = \max_w \frac{P(y, w|X)}{P(y|X)} \quad * \text{ when maximizing } \frac{P(y, w|X)}{P(y|X)}, \\ P(y, w|X) \text{ depends on } w.$$

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

$$\begin{aligned} \Rightarrow w^* &= \max_w P(y, w|X) \\ &= \max_w P(y|X, w)p(w) \rightarrow \text{From eq(A)} \end{aligned}$$

Applying log

$$\begin{aligned} w^* &= \max_w \log(P(y|X, w)p(w)) \\ &= \max_w \log P(y|X, w) + \log p(w) \\ &= \max_w \log C_1 \cdot \exp\left(-\frac{1}{2} (y - w^T X)^T [\sigma^2 I]^{-1} (y - w^T X)\right) \\ &\quad + \log C_2 \cdot \exp\left(-\frac{1}{2} w^T [\gamma^2 I]^{-1} w\right) \\ &= \max_w \left[-\frac{1}{2\sigma^2} (y - w^T X)^T (y - w^T X) - \frac{1}{2\gamma^2} w^T w \right] \\ &= \max_w \left[-\frac{1}{2\sigma^2} \|y - w^T X\|^2 - \frac{1}{2\gamma^2} \|w\|^2 \right] \end{aligned}$$

$$\begin{aligned} P(w|y, X) &= \frac{P(y, w|X)}{P(y|X)} \\ &= \frac{P(w|y|X)}{P(y|X)} \\ &= . \end{aligned}$$

Multiply by -1×20^2

(a)

$$= \min_{\omega} \underbrace{\|y - \omega^T x\|^2}_{\text{Sum of squares}} + \lambda \|\omega\|^2$$

λ $\|\omega\|^2$
L2-regularizer

$$\Rightarrow \omega^* = \max_{\omega} p(\omega | y, x) = \min_{\omega} \left[\|y - \omega^T x\|^2 + \lambda \|\omega\|^2 \right]$$

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

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

$$P(\theta | X) = \frac{\text{Likelihood} \cdot \text{Prior}}{P(X) \text{ evidence}}$$

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 $P(X)$

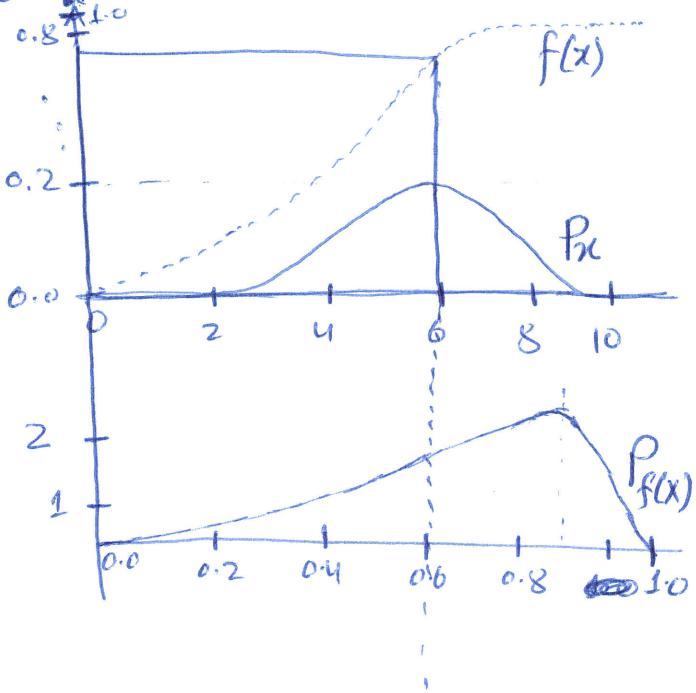
Thus it becomes an optimization problem.

$$\theta_{\text{MP}} = \underset{\theta}{\operatorname{argmax}} P(x|\theta) P(\theta)$$

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

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

Bayes formula:

$$P(\theta|X) = \frac{P(X|\theta) P(\theta)}{P(X)} \rightarrow \begin{array}{l} \text{Fixed by model} \\ \text{one own choice} \\ \text{Fixed by data} \end{array}$$

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

$$\xrightarrow{\mathcal{A}(v')} P(\theta|X) = \frac{P(X|\theta) P(\theta)}{P(X)} \downarrow \mathcal{A}(v)$$

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 "mean" μ & "variance" σ^2 then we would expect the posterior to also be normal but with some other mean & variance.

Example

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

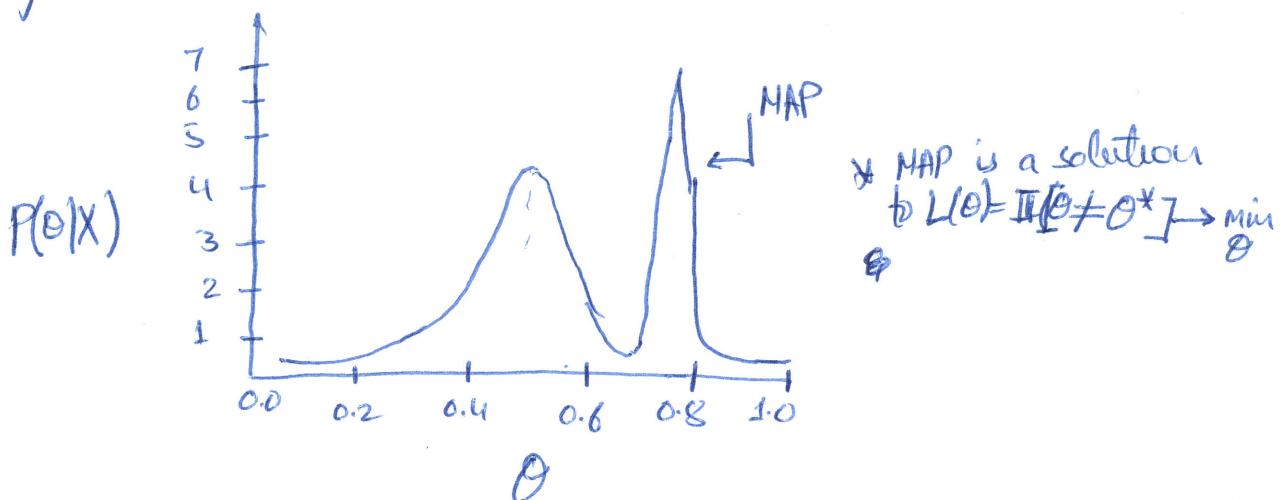
$$\mathcal{A}(v) = ??$$

$$\xrightarrow{\mathcal{A}(v')} P(\theta|X) = \frac{P(X|\theta) P(\theta)}{P(X)} \rightarrow N(\theta, \sigma^2) \downarrow \mathcal{A}(v)$$

what should be the conjugate prior $\mathcal{A}(v)$?

- A: If we select $\mathcal{A}(v)$ as a normal distribution then $P(\theta|X)$ will also have normal distribution. It is because product of two normal distributions $P(X|\theta) \cdot P(\theta)$ yields normal distribution $N(\theta|\bar{\theta}, \sigma^2)$
- i.e. $P(\theta|X) = \frac{P(X|\theta) P(\theta)}{P(X)} \leftarrow N(\theta|\bar{\theta}, \sigma^2)$
- $$\downarrow \mathcal{A}(v') N(\theta|a, b^2)$$

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



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. These would lead to the mean or median of the posterior distribution.

Objectives

$$L(\theta) = \prod [\theta \neq \theta^*] \rightarrow \min_{\theta}$$

$$U(\theta) = E(\theta - \theta^*)^2 \rightarrow \min_{\theta}$$

$$L(\theta) = E|\theta - \theta^*| \rightarrow \min_{\theta}$$

Solution

Mode

Mean

Median

And if you choose those 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 $D_{MP} = 12.53$ we can't say how confident are we.

Summarizing Maximum A posteriori Estimation

PROS

- * Easy to compute

CONS

- * Not invariant to reparametrization
- * Can't use as a prior
- * finds untypical point
- * Can't compute credible regions

* Another way to avoid computing evidence: conjugate distributions

Example

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

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

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

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 θ & variance σ^2 . Here $p(\theta)$ is not conjugate to $p(x|\theta)$.

W1-P2-L3

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 the axis)

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

Statistics

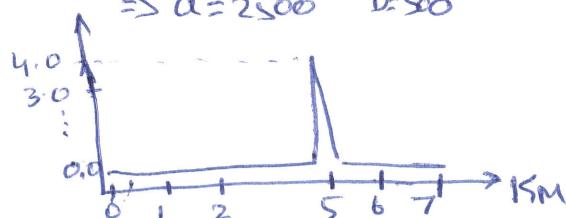
$$E[y] = ab \quad \underbrace{\text{mean}}_{\text{Mode}[y] = \frac{a-1}{b}} \quad \text{Var}[y] = ab^2$$

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

Example You run $5\text{ km} \pm 100\text{ m}$ a day. Model it by $\Gamma(y|a, b)$

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

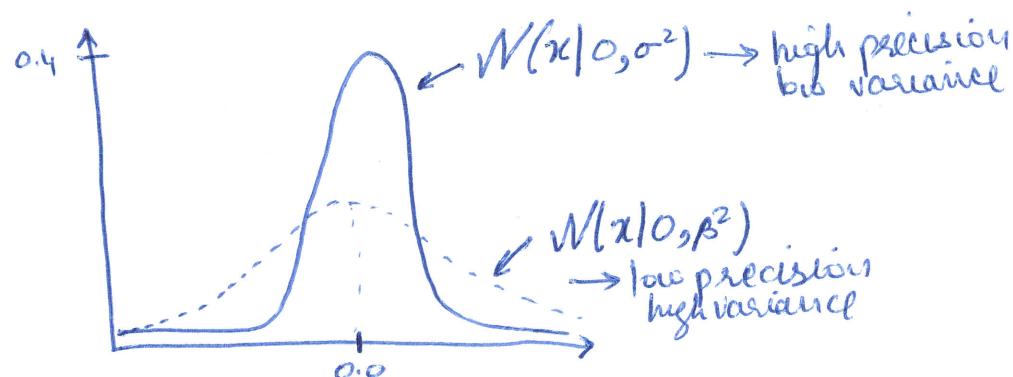
$$\Rightarrow a = 2500 \quad b = 500$$



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

$$\text{Precision} = \frac{1}{\text{Variance}} \quad (\gamma)$$

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



Precision

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

$$\text{As } \sigma^2 = 1/\gamma$$

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

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

Ans: Here is functional form

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

Dropping all constants

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

$$\propto \gamma^{1/2} e^{-\gamma(x-\mu)^2/2}$$

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

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

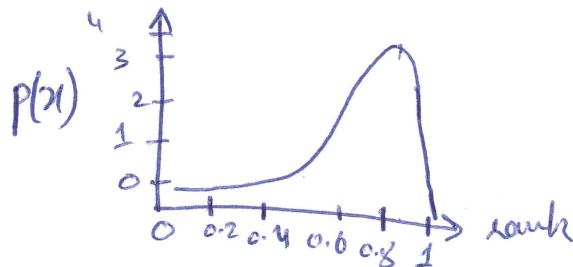
Support of beta distribution = $[0,1]$

Now: Use Beta Distribution to model our favorite movie rank

Movie rank is 0.8 ± 0.1

$$\text{E}[x] = \frac{a}{a+b} = 0.8$$

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



Actually, the Beta distribution is conjugate of the Bernoulli likelihood

$$P(X|\theta) = \theta^{N_1} (1-\theta)^{N_0} \rightarrow \text{Bernoulli likelihood} : \begin{matrix} N_1 \rightarrow \# \text{ of ones} \\ N_0 \rightarrow \# \text{ of zeros} \end{matrix}$$

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:

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

Use the Bayes formula:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto y e^{-y(b + \frac{(x-u)^2}{2})}$$

After dropping constants & rearranging terms

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

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

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

$$\text{where } N(x|u, y^{-1}) \neq \frac{\sqrt{y}}{\sqrt{2\pi}} e^{-\frac{(x-u)^2}{2}} \propto y^{\frac{1}{2}-by}$$

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

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

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

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

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

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

Dropping all constants yields:

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

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

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

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

Overall Picture

(17)

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

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
E.g $P(\theta|X) = B(N_1+a, N_0+b)$

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.

