



Bayesian Learning

Sergej Dogadov
s.dogadov@tu-berlin.de
Technische Universität Berlin - Machine Learning Group
Beginners Workshop Machine Learning 2019



Agenda

- First Part
 - Frequentist vs. Bayesian Approaches
 - Introduction to Probability Theory
 - Bayes' Theorem
 - Naive Bayes Classifier
 - Linear Regression
- Second Part
 - Bayesian Linear Regression
 - EM Algorithm
 - Variational Linear Regression
 - Variational Inference





Frequentist vs. Bayesian Approaches

Search for a missing base phone





Search for a missing base phone

Frequentist Approach

- Hear the phone beeping
- Identify an area only by the beeping sound
- Search for the phone

Bayesian Approach

- Hear the phone beeping
- Identify an area by the sound and the locations where I could misplace the phone in the past

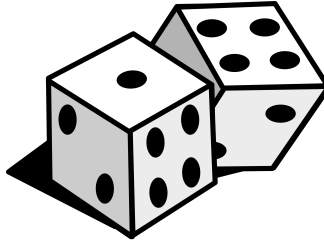
Bayesian approach incorporates the prior knowledge into the model in order to enhance the inference over the model's latent variables. In this case the location of the missing phone.

Example from: <https://stats.stackexchange.com/questions/22/bayesian-and-frequentist-reasoning-in-plain-english>





Introduction to Probability Theory





Probability Theory for reasoning under uncertainty

In many settings, we must try to understand what is going on in a system when we have imperfect or incomplete information.

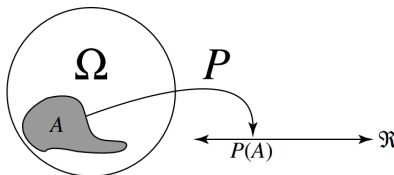
- Two reasons why we might reason under uncertainty:
 1. *laziness* (modeling every detail of a complex system is costly)
 2. *ignorance* (we may not completely understand the system)
- Probabilities quantify uncertainty regarding the occurrence of events.





Probability spaces

- A probability space represents our uncertainty regarding an experiment.
- It has two parts:
 1. the sample space Ω is a set of *outcomes* and
 2. the probability measure P , which is a real function of the subsets of Ω .



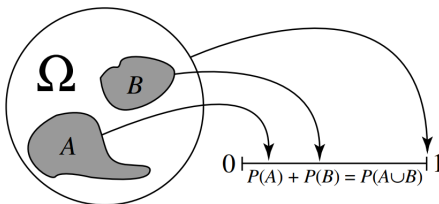
- A set of outcomes $A \subseteq \Omega$ is called an event. $P(A)$ represents how likely it is that the experiment's actual outcome will be a member of A .





The three axioms of Probability Theory

1. $P(A) \geq 0$ for all events A
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ for disjoint events A and B





Basic Formulas for Probabilities

Product Rule: probability $p(A \cap B)$ of a conjunction of two events A and B :

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

The probability that A and B both happen is the probability that A happens times the probability that B happens, given A has occurred.

Sum Rule: probability $p(A \cup B)$ of a disjunction of two events A and B :

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$





Bayes' Theorem

- **Bayes' Theorem** is a simple mathematical formula that has revolutionized how we understand and deal with uncertainty.
- It gives an answer to the question:

When we encounter new evidence, how much should it change our confidence in a belief?



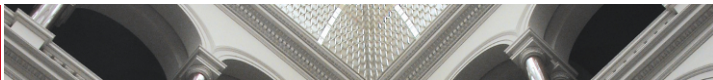


Bayes' Theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

- $p(A|B)$ is the probability of event A being true given that event B is true.
- $p(B|A)$ is the probability of event B being true given that event A is true.
- $p(A)$ and $p(B)$ are the probabilities of events A and B , where the events don't impact each other.

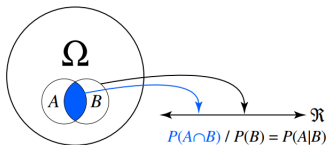




Bayes' Theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B \cap A)}{p(B)}$$

- $P(A)$ is called the *a priori* (or *prior*) probability of A and $P(A|B)$ is called the *a posteriori* probability of A given B .
- $p(B|A)$ is the **conditional** probability or **likelihood**, is the degree of belief in B , given that the proposition A is true.





Bayes' Theorem

In plain English, using Bayesian probability terminology, the Bayes rule can be written as

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

in other words

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

or

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$





Theorem of Total Probability

If events B_1, \dots, B_n are mutually exclusive with $\sum_{i=1}^n p(B_i) = 1$ (discrete case)

$$p(A) = \sum_{i=1}^n p(A|B_i)p(B_i)$$

and $\int_{\Omega} p(B)dB = 1$ (continuous case)

$$p(A) = \int_{\Omega} p(A|B)p(B)dB$$

also known as **marginalization**.





Theorem of Total Probability

Suppose that two factories supply light bulbs to the market. Factory X's bulbs work for over 5000 hours in 99% of cases, whereas factory Y's bulbs work for over 5000 hours in 95% of cases. It is known that factory X supplies 60% of the total bulbs available and Y supplies 40% of the total bulbs available.

What is the chance that a purchased bulb will work for longer than 5000 hours?

Applying the Theorem of Total Probability, we have:

$$P(A) = P(A|B_X)P(B_X) + P(A|B_Y)P(B_Y) = \quad (1)$$

(2)

$$\frac{99}{100} \cdot \frac{6}{10} + \frac{95}{100} \cdot \frac{4}{10} = \frac{594 + 380}{1000} = \frac{974}{1000} \quad (3)$$





Naïve Bayes Classifier

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$



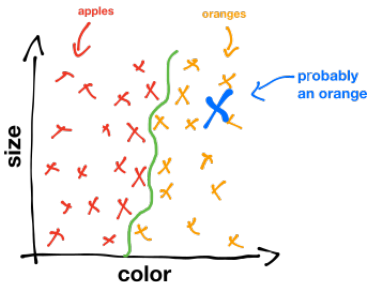
Thomas Bayes
1702 - 1761





Naive Bayes Classifier

Naive Bayes classifier is a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.





Probabilistic Model

Naive Bayes can be seen a conditional probability model, where each instance $\mathbf{x} = (x_1, \dots, x_n)$ with some n features needs to be classified.

The classifier assigns probabilities for each of K outcomes or **classes** C_k .

$$p(C_k|\mathbf{x}) = \frac{p(x_1, \dots, x_n|C_k)p(C_k)}{p(\mathbf{x})} \propto p(x_1, \dots, x_n, C_k)$$

The denominator does not depend on C and the values of the features x_i are given, so that the denominator is effectively constant.





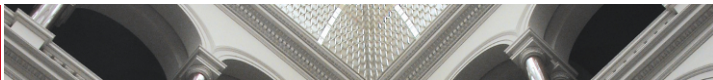
Probabilistic model

The "**naive**" assumption states that each feature x_i is conditionally independent of every other feature x_j for $i \neq j$, given the category C_k .

The joint model can be expressed as

$$p(C_k|\mathbf{x}) \propto p(x_1, \dots, x_n, C_k) = p(C_k)p(x_1|C_k)p(x_2|C_k) \cdots p(x_n|C_k) = \\ p(C_k) \prod_{i=1}^n p(x_i|C_k)$$





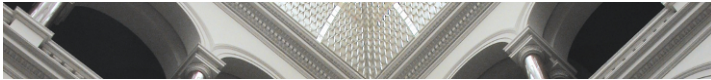
Naive Bayes Classifier

The naive Bayes classifier is the function that assigns a class label $\hat{y} = C_k$ for some k which is most probable.

The rule is also known as the **maximum a posteriori** (MAP).

$$\hat{y} = \operatorname{argmax}_{k \in \{1, \dots, K\}} p(C_k) \prod_{i=1}^n p(x_i | C_k)$$





Naive Bayes Classifier

Let's define a spam classifier using Bayes Theorem which can tell whether a given message is spam or not.





Naive Bayes Classifier

We have two classes $C_1 = \text{spam}$ and $C_2 = \text{ham}$. And each instance \mathbf{x} is a message with n words as features.

$$\mathbf{x} = (w_1, w_2, \dots, w_n)$$

In order to classify a message we need to determine which is greater

$$p(C = \text{spam} | w_1, w_2, \dots, w_n) \quad \text{or} \quad p(C = \text{ham} | w_1, w_2, \dots, w_n)$$





Naive Bayes Classifier

From the Bayes rule we know that the posterior can be expressed as

$$p(C = \text{spam} | w_1, w_2, \dots, w_n) \propto p(C = \text{spam}) \prod_{i=1}^n p(w_i | C = \text{spam})$$

Where

$$p(w_i | C = \text{spam}) = \frac{\text{Total number of occurrences of } w_i \text{ in spam messages}}{\text{Total number of words in spam messages}}$$

and

$$p(C = \text{spam}) = \frac{\text{Total number of spam messages}}{\text{Total number of messages}}.$$





Naive Bayes Classifier

How to deal with rare or non seen words ?

The 'trick' to avoid zero counts is so called *Laplace* smoothing.

Let V be the set of words in the training set.

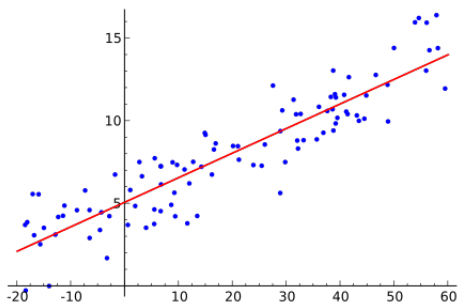
$$p(w_i | C = \text{spam}) = \frac{\text{Total number of occurrences of } w_i \text{ in spam messages} + 1}{\text{Total number of words in spam messages} + |V| + 1} \quad (4)$$





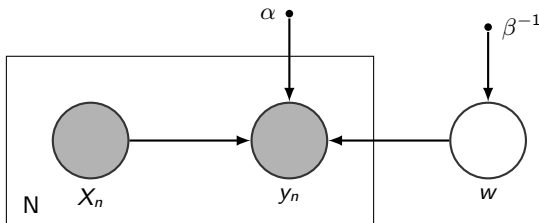
Bayesian Learning

Example of the Linear Regression model





Standard Linear Regression model (plate notation)



For each data point $x_i \in \mathbb{R}^{(1,k)}$, $i = \overline{1..N}$

$$y_i = x_i w + \epsilon_i$$

where $w \in \mathbb{R}^{(k,1)}$ is a regression vector and $\epsilon_i = \mathcal{N}(\epsilon_i|0, \alpha)$ is an independent zero-mean Gaussian noise with variance $\alpha \in \mathbb{R}^+$.





Prior of the regression vector

The prior over the regression vector w can be expressed with the precision parameter β as Gaussian with spherical covariance matrix:

$$p(w|\beta) = \mathcal{N}(w|\mu_0 = 0, \Sigma_0 = \beta^{-1}\mathbb{I}_k)$$

k -dimensional identity matrix:

$$\mathbb{I}_k = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_k \end{bmatrix}, \quad \text{where } d_i = 1, \quad i = \overline{1..k}$$





Multivariate Gaussian distribution

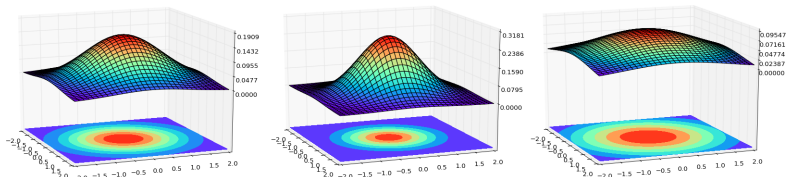


Figure: (left) $\beta = 1$, (middle) $\beta = 1.6$, (right) $\beta = 0.5$

PDF:

$$\mathcal{N}(x|\mu, \Sigma) = |\mathbf{2\pi\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

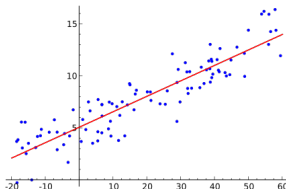
$$\ln \mathcal{N}(x|\mu, \Sigma) = -\frac{1}{2}\left(\ln 2\pi + \ln |\Sigma| + (x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$





Linear Regression generative model

Given $\mathbf{X} \in \mathbb{R}^{(N,k)}$ data points with $\mathcal{E} \sim \mathcal{N}(\mathcal{E}|0, \sigma^2 \mathbf{I}_N) \in \mathbb{R}^N$ noise.



Generative model for w as a given vector:

$$y = \mathbf{X}w + \mathcal{E} \quad \text{or} \quad y \sim \mathcal{N}(y|\mathbf{X}w, \sigma^2 \mathbf{I}_N) \in \mathbb{R}^N$$





Maximum a Posteriori

Task: to find a most probable value of w based on the observed data y and \mathbf{X} .

From Bayes rule we know that the posterior is proportional to likelihood times prior:

$$p(w|y, \mathbf{X}, \alpha, \beta) \propto p(y|\mathbf{X}, w, \alpha)p(w|\beta)$$

The log of the posterior can be written as following:

$$\mathcal{L}(w, \alpha, \beta) = \ln p(y|\mathbf{X}, w, \alpha) + \ln p(w|\beta) + \text{const}$$

$$\mathcal{L}(w, \alpha, \beta) = \ln \mathcal{N}(y|\mathbf{X}w, \alpha) + \ln \mathcal{N}(w|0, \beta^{-1}\mathbb{I}_k) + \text{const}$$





Apply log of the Gaussian distribution expression:

$$\mathcal{L}(w, \alpha, \beta) = -\frac{1}{2} \left(N \ln \alpha + \alpha^{-1} (y - \mathbf{X}w)^\top (y - \mathbf{X}w) - k \ln \beta + \beta w^\top w \right) + \text{const}$$

Maximum of the function for w can be found as following:

$$\frac{\partial \mathcal{L}(w, \alpha, \beta)}{\partial w} = \left(\alpha^{-1} \mathbf{X}^\top (y - \mathbf{X}w) - \beta w \right) = 0$$

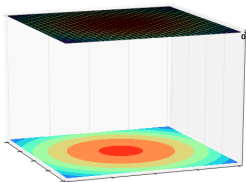
$$\alpha^{-1} \mathbf{X}^\top y = (\alpha^{-1} \mathbf{X}^\top \mathbf{X} + \beta \mathbb{I}_k) w$$

$$w^* = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I}_k \right)^{-1} \mathbf{X}^\top y, \quad \lambda = \alpha \beta$$





Broad prior



In case we put a broad prior ($\beta \rightarrow 0$) for the regression coefficients the *MAP* solution is

$$w^* = \mathbf{X}^\dagger y, \quad \text{where}$$

the quantity $\mathbf{X}^\dagger \equiv (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is known as **Moore-Penrose pseudo-inverse** of matrix $\mathbf{X} \in \mathbb{R}^{(N,k)}$.





Moore-Penrose pseudo-inverse

Properties:

$$\mathbf{X}\mathbf{X}^\dagger\mathbf{X} = \mathbf{X}$$

if there is an inverse of \mathbf{X} , then

$$\mathbf{X}\mathbf{X}^\dagger\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}\mathbf{X}^{-1} \Rightarrow \mathbf{X}\mathbf{X}^\dagger = \mathbb{I}$$

\mathbf{X}^\dagger is a *TRUE* inverse.

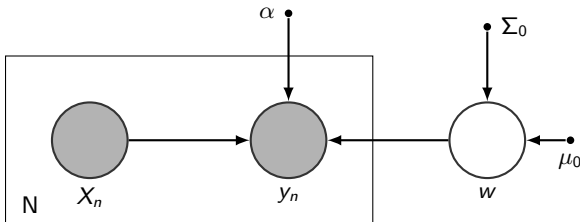




Linear Regression model with mean and covariance prior

Assume multivariate Gaussian prior for w with mean $\mu_0 \in \mathbb{R}^k$ and covariance matrix $\Sigma_0 \in \mathbb{R}^{(k,k)}$.

$$p(w|\mu_0, \Sigma_0) = \mathcal{N}(w|\mu_0, \Sigma_0)$$





Posterior distribution with mean and covariance

Task: to find a form of the posterior.

In this case the log posterior has the following form:

$$\mathcal{L}(w, \alpha, \mu_0, \Sigma_0) = \ln \mathcal{N}(y | \mathbf{X}w, \alpha) + \ln \mathcal{N}(w | \mu_0, \Sigma_0) + \text{const}$$

Apply log of the Gaussian distribution expression:

$$\mathcal{L}(w, \alpha, \mu_0, \Sigma_0) = -\frac{1}{2} \left(N \ln \alpha + \alpha^{-1} (y - \mathbf{X}w)^\top (y - \mathbf{X}w) + \right. \\ \left. \ln |\Sigma_0| + (w - \mu_0)^\top \Sigma_0^{-1} (w - \mu_0) \right) + \text{const}$$





Open brackets and consider the terms depending only on w .

$$\mathcal{L}(w) = -\frac{1}{2} \left(-2\alpha^{-1} w^\top \mathbf{X}^\top y + \alpha^{-1} w^\top \mathbf{X}^\top \mathbf{X} w - 2w^\top \mu_0 \Sigma_0^{-1} + w^\top \Sigma_0^{-1} w \right) + \text{const}$$

Combine terms within $w^\top \dots w$ and w^\top

$$\mathcal{L}(w) = -\frac{1}{2} \left(w^\top \overbrace{(\alpha^{-1} \mathbf{X}^\top \mathbf{X} + \Sigma_0^{-1})}^A w - 2w^\top \overbrace{(\alpha^{-1} \mathbf{X}^\top y + \Sigma_0^{-1} \mu_0)}^b \right) + \text{const}$$





Completing square

Mathematical trick useful for building quadric forms.

$$w^{\top}Aw - 2w^{\top}b = (w - A^{-1}b)^{\top}A(w - A^{-1}b) - b^{\top}A^{-1}b$$

Proof.

$$\begin{aligned}(w - A^{-1}b)^{\top}A(w - A^{-1}b) &= w^{\top}Aw - 2w^{\top}A^{-1}Ab + b^{\top}A^{-1}AA^{-1}b = \\ &= w^{\top}Aw - 2w^{\top}b + b^{\top}A^{-1}b\end{aligned}$$





Gaussian posterior distribution

After completing the square the posteriors has a form of Gaussian multivariate distribution with mean μ_N and Σ_N .

$$p(w|y, \mathbf{X}, \alpha, \mu_0, \Sigma_0) = \mathcal{N}(w|\mu_N, \Sigma_N)$$

$$\Sigma_N = \left(\alpha^{-1} \mathbf{X}^T \mathbf{X} + \Sigma_0^{-1} \right)^{-1}$$

$$\mu_N = \Sigma_N (\alpha^{-1} \mathbf{X}^T y + \Sigma_0^{-1} \mu_0)$$

or

$$\Sigma_N = \left(\mathbf{X}^T \mathbf{X} + \alpha \Sigma_0^{-1} \right)^{-1}$$

$$\mu_N = \Sigma_N (\mathbf{X}^T y + \alpha \Sigma_0^{-1} \mu_0)$$





Properties of the Gaussian Posterior

- Because the posterior is Gaussian its mode coincides with its mean. Thus the maximum posterior weight is given by $w_{MAP} = \mu_N$
- If we consider infinitely broad prior $\Sigma_0 = \beta^{-1}\mathbb{I}$, $\beta \rightarrow 0$, the mean of the posterior reduces to the MAP solution without regularization term.

$$\mu_N = \mathbf{X}^\dagger \mathbf{y}$$

- If $N = 0$, posterior reverts to the prior.
- If data points arrive sequentially, then posterior to any stage acts as prior distribution for subsequent data points.





Sequential Bayesian Learning example

We can illustrate the sequential update of a posterior in a simple example.

- Consider a linear model of the form $y(x, a) = a_0 + a_1x$ where x is a single input variable and $a_0 = -0.3$ and $a_1 = 0.5$
- We can generate the synthetic data by choosing the values of x from univariate distribution $x_n \sim \mathcal{U}(x | -1, 1)$ then evaluate $y(x_n, w)$
- Finally add some Gaussian noise with standard deviation $\alpha = 0.2$ to obtain the target value y_n

Example from: Christopher M. Bishop "Pattern Recognition and Machine Learning" P. 155



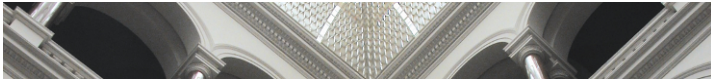


Illustration of sequential Bayesian I

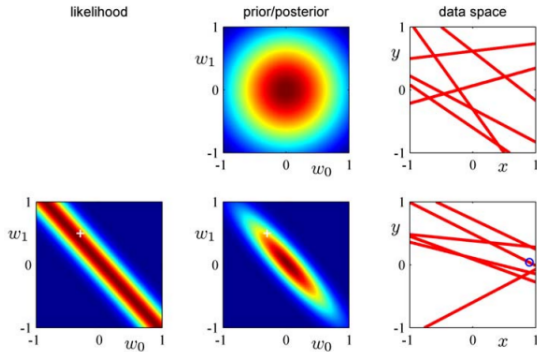
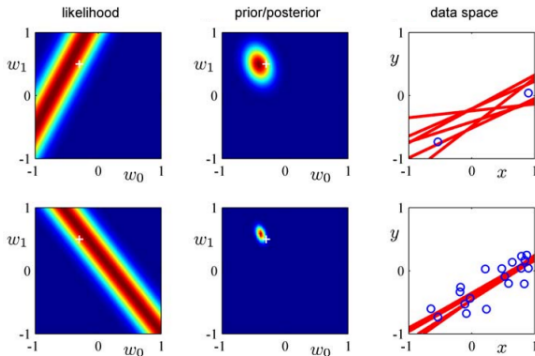
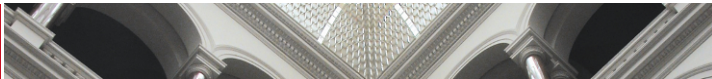




Illustration of sequential Bayesian II





End of the first part.





Expectation Maximization motivation

Question: How to optimize the model parameters ?

Find the maximum of the log-likelihood function w.r.t the model parameters and the expected value of w .

Update the model variable and repeat the procedure till the convergence
(*Expectation maximization* algorithm)

$$\mathbb{E}[\mathcal{L}(w, \alpha, \mu_0, \Sigma_0)] = -\frac{1}{2} \left(N \ln \alpha + \alpha^{-1} \mathbb{E} \left[(y - \mathbf{X}w)^\top (y - \mathbf{X}w) \right] \right. \\ \left. + \ln |\Sigma_0| + \mathbb{E} \left[(w - \mu_0)^\top \Sigma_0^{-1} (w - \mu_0) \right] \right) + \text{const}$$





Some math...

Consider terms depending on the noise factor α .

$$\mathcal{L}(\alpha) = -\frac{1}{2} \left(N \ln \alpha + \alpha^{-1} \mathbb{E} \left[(y - \mathbf{X}w)^\top (y - \mathbf{X}w) \right] \right) + \text{const}$$

$$\frac{\partial \mathcal{L}(\alpha)}{\partial \alpha} = -\frac{1}{2} \left(N \alpha^{-1} - \alpha^{-2} \mathbb{E} \left[(y - \mathbf{X}w)^\top (y - \mathbf{X}w) \right] \right) = 0$$

$$N \alpha^{-1} = \alpha^{-2} \mathbb{E} \left[(y - \mathbf{X}w)^\top (y - \mathbf{X}w) \right] \quad | \quad * \alpha^2$$

$$\hat{\alpha} = \frac{y^\top y - 2y^\top \mathbf{X} \mathbb{E}[w] + \text{Tr} \left(\mathbb{E}[ww^\top] \mathbf{X}^\top \mathbf{X} \right)}{N}$$





Noise variance optimization α

Because the posterior is Gaussian with mean μ_N and covariance Σ_N the first and the second moment are

$$\mathbb{E}[w] = \mu_N, \quad \mathbb{E}[ww^\top] = \mu_N \mu_N^\top + \Sigma_N$$

The updated noise variance value is then

$$\hat{\alpha} = \frac{y^\top y - 2y^\top \mathbf{X} \mu_N + \text{Tr}\left((\mu_N \mu_N^\top + \Sigma_N) \mathbf{X}^\top \mathbf{X}\right)}{N}$$





Mean of the prior distribution optimization μ_0

Consider terms depending on μ_0 .

$$\mathbb{E}[\mathcal{L}(\mu_0, \Sigma_0)] = -\frac{1}{2} \left(\ln |\Sigma_0| + \mathbb{E} \left[(w - \mu_0)^\top \Sigma_0^{-1} (w - \mu_0) \right] \right) + \text{const}$$

$$\frac{\partial \mathbb{E}[\mathcal{L}(\mu_0)]}{\partial \mu_0} = -\mu_0 \Sigma_0^{-1} + \mathbb{E}[w] \Sigma_0^{-1} = 0$$

$$\hat{\mu}_0 = \mathbb{E}[w] = \mu_N$$





Covariance of the prior distribution optimization Σ_0

Consider terms depending on Σ_0 .

$$\frac{\partial \mathbb{E}[\mathcal{L}(\Sigma_0)]}{\partial \Sigma_0} = \Sigma_0^{-1} - \left(\mathbb{E}[\mathbf{w}\mathbf{w}^\top] - 2\mathbb{E}[\mathbf{w}]\hat{\mu}_0 + \hat{\mu}_0\hat{\mu}_0^\top \right) \Sigma_0^{-2} = 0$$

$$\hat{\Sigma}_0 = \mathbb{E}[\mathbf{w}\mathbf{w}^\top] - 2\mathbb{E}[\mathbf{w}]\hat{\mu}_0 + \hat{\mu}_0\hat{\mu}_0^\top$$

$$\hat{\Sigma}_0 = \mu_N \mu_N^\top + \Sigma_N - 2\mu_N \mu_N^\top + \mu_N \mu_N^\top = \Sigma_N$$

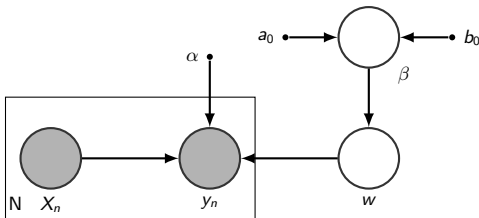
$$\boxed{\hat{\Sigma}_0 = \Sigma_N}$$





Fully Bayesian Linear model

Consider a prior distribution not only over the regression vector w but also over its precision parameter β .



$$p(\beta|a_0, b_0) = \mathcal{G}(\beta|a_0, b_0) \quad (\text{Gamma distribution})$$

The joint distribution of all variables is given by:

$$p(y, \mathbf{w}, \beta | \mathbf{X}, \alpha) = p(y | \mathbf{w}) p(\mathbf{w} | \beta) p(\beta)$$

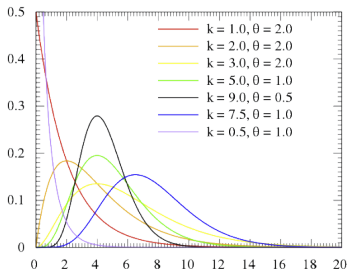




Gamma distribution

$$\mathcal{G}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp^{-bx}$$

Gamma PDF with $k = a$ and $\theta = 1/b$



$$\mathbb{E}[x] = \frac{a}{b}, \quad \mathbb{E}[\ln x] = \psi(a) - \ln b$$





Non tractable posterior

The joint distribution of all variables is

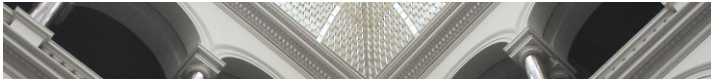
$$p(y, w, \beta | \mathbf{X}, \alpha) = \mathcal{N}(y | \mathbf{X}w, \alpha) \mathcal{N}_k(w | 0, \beta^{-1} \mathbb{I}_k) \mathcal{G}(\beta | a_0, b_0)$$

The posterior distribution $q(w, \beta | y, \mathbf{X}, a, b, \alpha)$ for the fully Bayesian Linear model unfortunately has no closed form solution and therefore intractable.

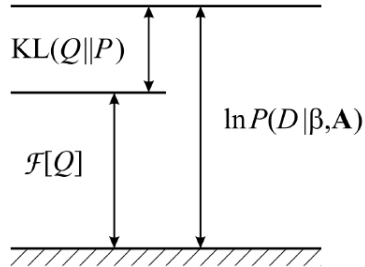
Goal: to find an approximation to the posterior $p(w, \beta | y, \mathbf{X}, a, b, \alpha)$.

In order to find the approximation we can employ the *Variational Inference* framework.





Variational Inference





Suppose we have a fully Bayesian model in which all parameters are given prior distributions.

- The set of all observed variables denoted by \mathbf{X} and hidden variables by \mathbf{Z} .
- Our probabilistic model specifies the joint distribution $p(\mathbf{X}, \mathbf{Z})$, and our goal is to find an approximation for the posterior distribution $p(\mathbf{Z}|\mathbf{X})$.

Similar to EM Algorithm we can decompose the log marginal probability $p(\mathbf{X})$ as following:

$$p(\mathbf{X}) = \mathcal{L}(q) + \mathbb{D}_{KL}(p||q)$$

This differs from our discussion of EM only in that the parameter vector θ no longer appears, because the parameters are now stochastic variables and are absorbed into \mathbf{Z} .





Factorized distributions

The EM Algorithm cannot be applied because the form of the posterior distribution is intractable.

Solution: to partition the elements of \mathbf{Z} into disjoint groups that we denote by \mathbf{Z}_i where $i = 1, \dots, M$ so that

$$q(\mathbf{Z}) = \prod_{i=1}^M q(\mathbf{Z}_i) \quad \text{mean field assumption}$$

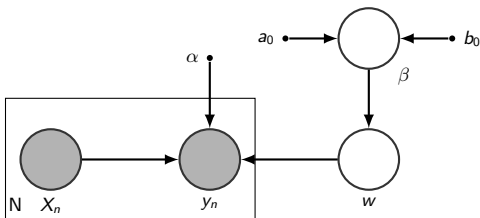
Optimal solutions for the approximated posterior has following form:

$$\ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j} \left[\ln p(\mathbf{X}, \mathbf{Z}) \right] + \text{const}$$





Fully Bayesian Linear model solution



The joint distribution of all variables is

$$p(y, \mathbf{w}, \beta | \mathbf{X}, \alpha, a_0, b_0) = p(y | \mathbf{w}, \mathbf{X}, \alpha) p(\mathbf{w} | \beta) p(\beta | a_0, b_0)$$

$$q^*(\beta) = \mathbb{E}_w [\ln p(\mathbf{w} | \beta)] + \ln p(\beta | a_0, b_0) + \text{const}$$

and

$$q^*(\mathbf{w}) = \ln p(y | \mathbf{w}, \mathbf{X}, \alpha) + \ln \mathbb{E}_\beta [p(\mathbf{w} | \beta)] + \text{const}$$





Approximate posterior for precision β

$$\ln q^*(\beta) = \mathbb{E}_w[\ln \mathcal{N}_k(w|0, \beta^{-1}\mathbb{I}_k)] + \ln \mathcal{G}(\beta|a_0, b_0) + \text{const}$$

$$\ln q^*(\beta) = \frac{k}{2} \ln \beta - \frac{\beta}{2} \mathbb{E}[w^\top w] + (a_0 - 1) \ln \beta - b_0 \beta + \text{const}$$

$$\ln q^*(\beta) = (a_0 + \frac{k}{2} - 1) \ln \beta - (b_0 + \frac{\mathbb{E}[w^\top w]}{2}) \beta + \text{const}$$

$$q^*(\beta) = \mathcal{G}(\beta|a_N, b_N), \quad \text{where}$$

$$a_N = a_0 + \frac{k}{2}, \quad b_N = b_0 + \frac{\mathbb{E}[w^\top w]}{2}$$





Approximate posterior for weight vector w

$$q^*(w) = \mathcal{N}_k(w | \mu_N, \Sigma_N), \quad \text{where}$$

$$\mu_N = \Sigma_N \mathbf{X}^\top y, \quad \Sigma_N = \left(\alpha \mathbb{E}[\beta] \mathbb{I}_k + \mathbf{X}^\top \mathbf{X} \right)^{-1}$$

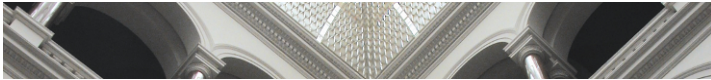
First moment of the Gamma distribution:

$$\mathbb{E}[\beta] = \frac{a_N}{b_N}$$

Knowing the second moment of the Gaussian distribution we have following:

$$\mathbb{E}[w^\top w] = \text{Tr}(\mathbb{E}[ww^\top]) = \text{Tr}(\mu_N \mu_N^\top + \Sigma_N) = \mu_N^\top \mu_N + \text{Tr}(\Sigma_N)$$





THANK YOU!

