Probabilistic Machine Learning:
1. Probabilistic refresher

Tomasz Kajdanowicz, Piotr Bielak, Maciej Falkiewicz, Kacper Kania, Piotr Zieliński

Department of Computational Intelligence
Wroclaw University of Science and Technology

The presentation has been inspired and in some parts totally based on

1. Prof. Mario A. T. Figueiredo presentation at LxMLS'2017, Instituto Superior Tecnico & Instituto de Telecomunicacoes, Lisboa, Portugal
2. working notes of Prof. Martin Ridout,University of Kent, UK
3. Murphy's book
4. working notes of Thomas Schoon, Uppsala University
5. working notes of Larry Wasserman, Machine Learning Department, Carnegie Mellon University
6. working notes of Tom Loredo, Deptartment of Astronomy, Cornell University
7. https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7

Table of Contents

Pre-reading

1. Random Variables
   - 1.1 Probability Functions and Distribution Functions
   - 1.2 Continuous Random Variables
   - 1.3 Functions of Random Variables
   - 1.4 Joint Distributions
   - 1.5 Independence of Random Variables
   - 1.6 Conditional Distributions
2. Expectation, Variance, Moments
   - 2.1 Expected Value
   - 2.2 Variance and Standard Deviation
   - 2.3 Moments and Generating Functions
   - 2.4 Covariance and Correlation
   - 2.5 Conditional Expectation
   - 2.6 Median and Quantiles

Table of Contents

Bayesian probability

In Bayesian view probabilities provide a quantification of **uncertainty**.

Example

- lets have uncertain event that can not be repeated numerous times in order to define a notion of probability, e.g. whether the Arctic ice cap will have disappeared by the end of the century
- we have some general idea, e.g. of how quickly we think the polar ice is melting
- having obtained fresh evidence, (satellite Earth images) we may revise our opinion on the rate of ice loss
- we would like to be able to quantify our expression of uncertainty and make precise revisions of uncertainty in the light of new evidence

Bayesian probability is an *interpretation* of the concept of probability

Instead of frequency of some phenomenon, probability is interpreted as:
reasonable expectation, representing a state of knowledge or quantification of a personal belief

Bayesian probability

Bayes evaluation of the uncertainty

If we have a general idea on the model of phenomenon and w are parameters of that model:

- we capture our assumptions about w, before observing the data, in the form of a **prior probability distribution** $p(w)$
- the effect of the observed data $D = \{t_1, ..., t_N\}$ is expressed through the **conditional probability** $p(D|w)$

Bayes' theorem

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

allows us to evaluate the uncertainty in w after we have observed D in the form of posterior probability $p(w|D)$

Likelihood function and Bayes theorem

- $p(w|D)$ is evaluated for the observed dataset D
- can be viewed as a function of the parameter vector w
- is called **likelihood function**
- express how probable the observed data is for different settings of the parameters w
- likelihood is not a probability distribution over w (its integral with respect to w does not (necessarily) equal one
-

Bayes' theorem in words

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

where all of these quantities are viewed as functions of w

The denominator in s the normalization constant, which ensures that the posterior distribution

Comments on Bayes theorem

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

- the denominator is the normalization constant, which ensures that the posterior distribution is valid probability density and integrates to one
- integrating with respect to w the denominator in terms of the prior distribution and the likelihood function is

$$p(D) = \int p(D|w)p(w)dw$$

Frequentist setting

- w is fixed
- value is determined by some 'estimator'
- error bars on this estimate are obtained by considering the distribution of possible data sets D

Bayesian viewpoint

- there is only single dataset D
- the uncertainty in the parameters is expressed through a probability distribution over w

Why we analyze data?

- Scientists **argue!**
- Argument $\equiv$ Collection of statements comprising an act of reasoning from premises to a conclusion
- **A key goal of science:** Explain or predict quantitative measurements (data!)
- **Data analysis:** Constructing and appraising arguments that reason from data to interesting scientific conclusions (explanations, predictions)

The role of data

Remark

Data do not speak for themselves!

- We don't just tabulate data, we analyze data
- We gather data so they may speak for or against existing hypotheses, and guide the formation of new hypotheses
- A key role of data in science is to be among the premises in scientific arguments

More on Frequentist vs. Bayesian statements

"The data D support conclusion C . . ."

Frequentist assessment

"C was selected with a procedure that is right 95% of the time over a set $\{D_{hyp}\}$ that includes D"

Probabilities are properties of procedures, not of particular results.

Bayesian assessment

"The strength of the chain of reasoning from the model and D to C is 0.95, on a scale where 1= certainty."

Probabilities are associated with specific, observed data.

Law of Total Probability (LTP)

- fundamental rule relating marginal probabilities to conditional probabilities

LTP

If $\{B_i\}$ for $i = 1, 2, 3, \ldots$ is exclusive, exhaustive and at least one of them is true

$$P(A) = \sum_i P(A, B_i)$$

or alternatively

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

Example of TLP

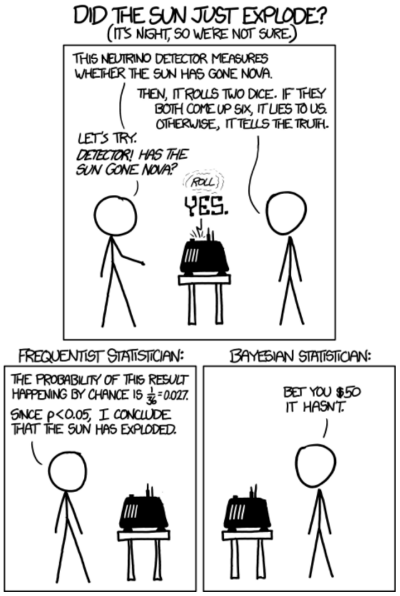
- two factories supply light bulbs
  - factory X's bulbs work for over 5000 hours in 99% of cases, supplies 60% of the total bulbs available (B_X)
  - factory Y's bulbs work for over 5000 hours in 95% of cases, supplies 40% of the total bulbs available (B_Y)
- what is the chance that a purchased bulb will work for longer than 5000 hours (A)?

Calculations

$$\begin{aligned} P(A) &= P(A|B_X)P(B_X) + P(A|B_Y)P(B_Y) \\ &= \frac{99}{100}\frac{6}{10} + \frac{95}{100}\frac{4}{10} = 97.4\% \end{aligned}$$

Table of Contents

Who sees the difference?



Bayesian versus Frequentist Inference

	Bayesian	Frequentist
Probability	subjective degree of belief	limiting frequency
Goal	analyze beliefs	create procedures with frequency guarantees
Model parameters Θ	random variable	fixed
Input X	random variable	random variable
Use Bayes' theorem?	Yes. To update beliefs.	Yes, if it leads to procedure with good frequentist behavior. Otherwise no.

Bayesian inference mechanics

Lets have:

- x - a data point
  (can be a vector of values)
- θ - parameter of the data point's distribution, $x \sim p(x \mid \theta)$
  (can be a vector of parameters)
- α - hyperparameter of the parameter distribution, i.e., $\theta \sim p(\theta \mid \alpha)$
  (can be a vector of hyperparameters)
- D is the sample, a set of n observed data points, i.e. $x_1, \ldots, x_n$
- $\tilde{x}$, a new data point whose distribution is to be predicted

Bayesian inference mechanics

GOAL: Infere θ parameters conditioned on data

- **prior distribution:** $p(\theta \mid \alpha)$
  distribution of the parameter(s) before any data is observed
- **likelihood** (sampling distribution): $p(D \mid \theta)$
  distribution of the observed data conditional on its parameters, a function of the parameters $\mathrm{L}(\theta \mid D) = p(D \mid \theta)$
- **marginal likelihood** (evidence): $p(D \mid \alpha) = \int p(D \mid \theta) p(\theta \mid \alpha) \, \mathrm{d}\theta$
  distribution of the observed data marginalized over the parameter(s)
- **posterior distribution**:
  distribution of the parameter(s) after taking into account the observed dataset

$$p(\theta \mid D, \alpha) = \frac{p(\theta, D, \alpha)}{p(D, \alpha)} = \frac{p(D \mid \theta, \alpha) p(\theta, \alpha)}{p(D \mid \alpha) p(\alpha)} = \frac{p(D \mid \theta, \alpha) p(\theta \mid \alpha)}{p(D \mid \alpha)} \propto p(D \mid \theta, \alpha) p(\theta \mid \alpha)$$

Where Does the Prior Come From?

- a million dollar question
- it is supposed to choose a prior that represents their prior information
- challenging in high dimensional cases (prior ends up being highly influential)
- may try noninformative priors, e.g. *Jeffreys prior*
- it is common to use flat priors

Empirical Bayes

How do we decide on the suitable values for hyperparameters α?
IDEA: Estimate the hyperparameters from the data by selecting them such that they maximize the marginal likelihood function

Empirical Bayes

Empirical Bayes combines the two statistical philosophies:

- frequentistic ideas are used to estimate the hyperparameters
- then used within the Bayesian inference

Other names of empirical Bayes: type 2 maximum likelihood, generalized maximum likelihood, and evidence approximation.

Large Sample Theory

There is a Bayesian central limit theorem.For models, with large n (samples)

$$p(\theta \mid D) \approx \mathcal{N}\left(\hat{\theta}, \frac{1}{f(\hat{\theta})}\right)$$

where

- $\hat{\theta}$ is the MLE
- f is Fisher information (way of measuring the amount of information that an observable random variable carries about an unknown parameter)

Conclusion on Bayesian Inference

Frequentist inference answers the question:
How do I construct a procedure that has frequency guarantees?

Bayesian inference answers the question:
How do I update my subjective beliefs after I observe some data?

Table of Contents

The **number game**

Rules:

- I choose some simple arithmetical concept C, such as "prime number" or "a number between 1 and 10"
- I then give you a series of randomly chosen positive examples $D = \{x_1, ..., x_N\}$ drawn from C, and ask you whether some new test case $\tilde{x}$ belongs to C, I ask you to classify $\tilde{x}$
- for simplicity all numbers are integers between 1 and 100

The **number game**

- what if I tell you "16" is a positive example of the concept?
- what other numbers do you think are positive? 17? 6? 32? 99? It's hard to tell with only one example.
- presumably numbers that are similar in some sense to 16 are more likely, but similar in what way?
- but some numbers are more likely than others!
- **posterior predictive distribution**:
  $p(\tilde{x}|D)$, which is the probability that $\tilde{x} \in C$ given the data D for any $\tilde{x} \in \{1, \ldots, 100\}$

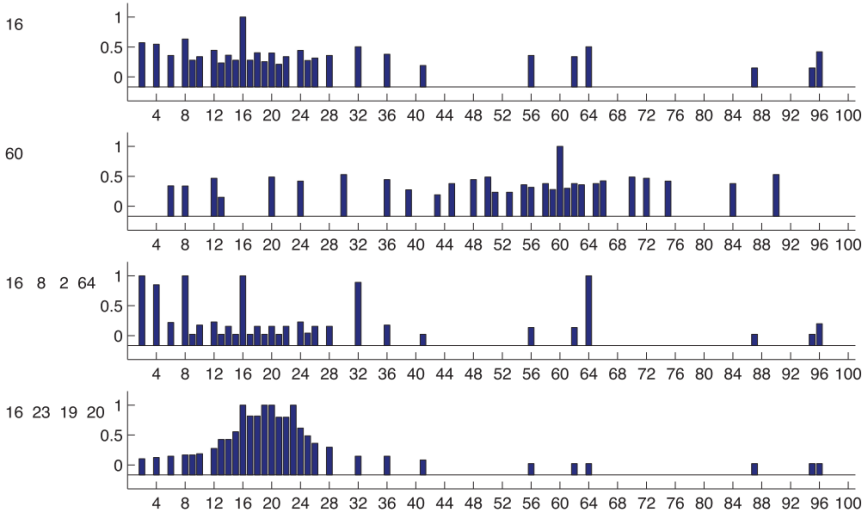
The **number game**

- what if I tell you "16", "8", "2" and "64" are also positive examples of the concept?
- you say: the concept is "powers of two"
- this is an example of **induction**

- Tell, what is the concept for D=16?

- what if I tell you the data is $D = \{16, 23, 19, 20\}$?

The **number game**



How to capture it in machine?

- classic approach: **induction** - suppose a hypothesis space of concepts H (e.g. odd numbers, even numbers, all numbers between 1 and 100, powers of two, all numbers ending in j (for $0 \leq j \leq 9$), etc.)
- **version space**: subset of H that is consistent with the data D
- BUT after seeing $D = \{16\}$, there are many consistent rules, how to combine them to predict if $\tilde{x} \in C$
- good explanation: the Bayesian explanation

Likelihood

- explain why after seeing $D = \{16, 8, 2, 64\}$ we choose $h_{two} \doteq$ "powers of two", and not, say, $h_{even} \doteq$ "even numbers"
- lets assume that examples are sampled uniformly at random from the **extension of a concept** (set of numbers that belong to it)
- then the probability of independently sampling N items (with replacement) from h is

$$p(D|h) = \left(\frac{1}{size(h)}\right)^N = \left(\frac{1}{|h|}\right)^N$$

- REMARK: favors the simplest (smallest) hypothesis consistent with the data: **Occam's razor**

Wiliam Occam

How likelihood works

- let $D = \{16, 8, 2, 64\}$
- then $p(D|h_{two}) = \frac{1}{6}$ (only 6 powers of two less than 100)

  $p(D|h_{even}) = \frac{1}{50}$ (50 even numbers)

- after 4 examples (N=4)

  $h_{two} = (\frac{1}{6})^4 = 0.00077$

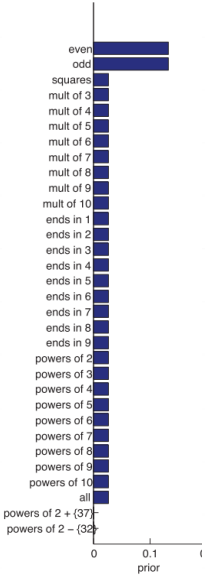
  $h_{even} = (\frac{1}{50})^4 = 0.00000016$

Prior

- let $D = \{16\}$
- $h' \doteq$ "powers of two except 32" is more likely than $h \doteq$ "powers of two"
- h' conceptually unnatural
- to capture such intuition: assign low prior probability to unnatural concepts
- is **subjective**

- in our example lets use a simple prior:
- uniform probability on 32 simple arithmetical concepts:
  - (1) even numbers, (2) odd numbers
  - (3) squares
  - (4) - (11) multiplication
  - (12) - (20) "numbers ending in $a \in \{1, 2, \ldots, 9\}$"
  - (21) - (29) power of 2, of 3, ..., 10
  - (30) all
  - (31) - (32) "powers of 2, plus 37", "powers of 2, except 32" (unnatural, low prior weight)

Prior in the number game



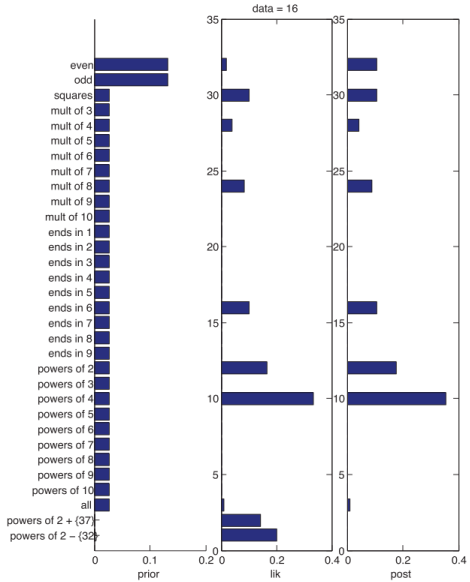
Posterior

- simply: posterior is the likelihood times the prior, normalized

$$p(h|D) = \frac{p(D|h)p(h)}{\sum_{h' \in H} p(D, h')} = \frac{p(h)1(D \in h)/|h|^N}{\sum_{h' \in H} p(h')1(D \in h)/|h'|^N}$$

$1(D \in h)$ - 1 (true) if $D \in h$, otherwise 0

Prior, Likelihood, Posterior in the number game



Prior, Likelihood, Posterior in the number game

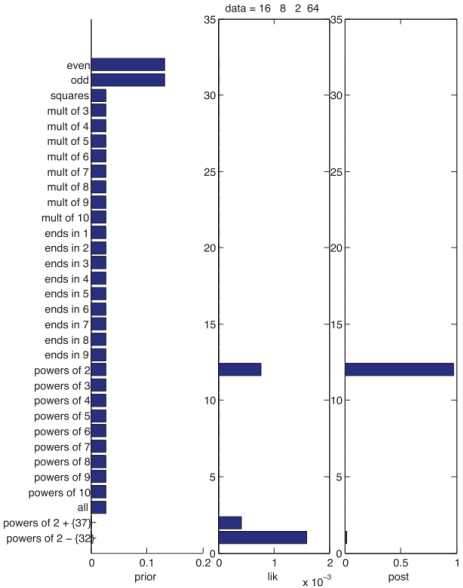


Table of Contents

Beta-binomial model

"Number game"

- inferring a distribution over a discrete variable drawn from a finite hypothesis space
- given a series of discrete observations
- computations particularly simple: sum, multiplication and division

What if, like in many applications, the unknown parameters are continuous?

- the hypothesis space is subset of $\mathbb{R}^K$, where K is the number of parameters
- replace sums with integrals

Coin toss example

The problem:

- inferring the probability that a coin shows up heads
- given a series of observed coin tosses

Might seem trivial, but

- this model forms the basis of many of the methods
- historically important, since it was the example which was analyzed in Bayes' original paper of 1763

Recipe of specifying the model

Define

- likelihood
- prior

and derive

- posterior
- posterior predictive

The problem

Let's consider a single binary random variable:

- $X_i \sim Bern(\theta)$
- $X_i = 1$ represents "heads", $X_i = 0$ represents "tails"
- $\theta \in [0, 1]$ is the parameter (probability of heads)
- $p(X_i = 1|\theta) = \theta, \;\; p(X_i = 0|\theta) = 1 - \theta$

Probability distribution over X

- $Bern(X|\theta) = \theta^{N_1}(1 - \theta)^{N_0}$
- if the data are iid, the likelihood has the same shape
- there are $N_1 = \sum_{i=1}^{N} 1(X_i = 1)$ heads and $N_0 = \sum_{i=1}^{N} 1(X_i = 0)$ tails
- N_0 and N_1 are called **sufficient statistics** (this is **all** we need to know about data to infer θ)

Bernoulli distribution recap

$Bern(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$
$N = N_0 + N_1$

Mean:

- $\mathbb{E}(X) = \theta$

Variance:

- $var(X) = \theta(1-\theta)$

The problem: continuing

Let's demistyfy the examplary problem more:

- suppose the data consists of the count of the number of heads N_1 observed in a fixed number $N = N_1 + N_0$ of trials
- $N_1 \sim Bin(N, \theta)$
- binomial pmf: $Bin(k|n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$
- term $\binom{n}{k}$ is constant independent of θ, thus the **binomial sampling model is the same as the likelihood for the Bernoulli model**

Likelihood: in general

Likelihood:

- a tool for summarizing the data's evidence about unknown parameter in the model
- as below: considered as a function of θ
- or: is the likelihood function (of θ)
- the probability of "the value x of X for the parameter value θ"

Discrete probability distribution

$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x)$

Continuous probability distribution

$\mathcal{L}(\theta \mid x) = f_\theta(x)$

Likelihood of our problem

$$\mathcal{L}(\mathcal{D} \mid \theta) = \theta^{N_1}(1-\theta)^{N_0}$$

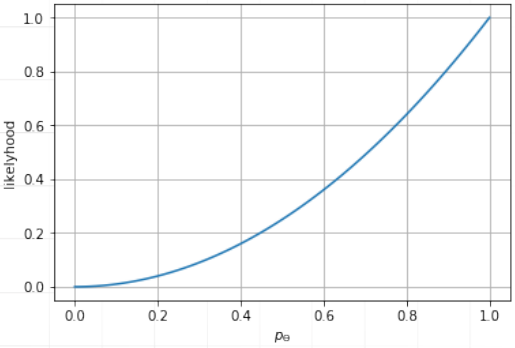


Figure: Likelihood $\mathcal{D}$={HH}

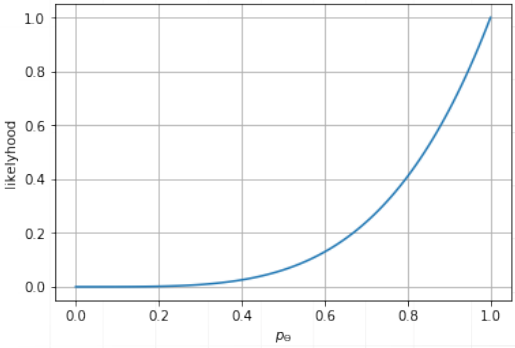


Figure: Likelihood $\mathcal{D}$={HHHH}

What about prior?

- need of prior with support over [0,1] interval
- easier, if it would have the same form as likelihood, for some prior parameters γ_1 and γ_2:

  $p(\theta) \propto \theta^{\gamma_1}(1-\theta)^{\gamma_2}$
- easy evaluation of posterior: adding exponents

$$\mathcal{L}(\theta \mid \mathcal{D})p(\theta) = \theta^{N_1}(1-\theta)^{N_0}\theta^{\gamma_1}(1-\theta)^{\gamma_2} = \theta^{N_1+\gamma_1}(1-\theta)^{N_0+\gamma_2}$$

Conjugate priors

When the prior and the posterior have the same form, we say that the prior is a conjugate prior for the corresponding likelihood.

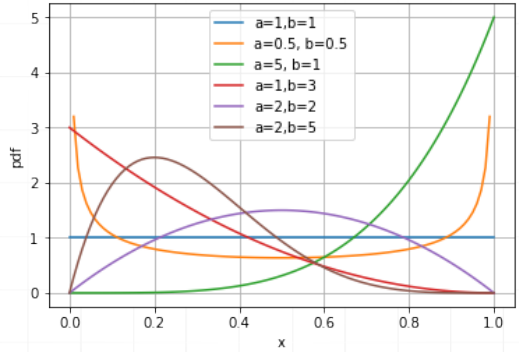
Conjugate priors

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
Binomial	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$\mathrm{BetaBin}(\tilde{x} \mid \alpha', \beta')$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures[note 1] (i.e., $\frac{\beta - 1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	k, θ	$k + \sum_{i=1}^{n} x_i,\ \frac{\theta}{n\theta + 1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\mathrm{NB}(\tilde{x} \mid k', \theta')$ (negative binomial)
			α, β[note 3]	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n$	α total occurrences in β intervals	$\mathrm{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1 + \beta'}\right)$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ occurrences of category i[note 1]	$p(\tilde{x} = i) = \frac{\alpha_i'}{\sum_i \alpha_i'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + \sum_{i=1}^{n} \boldsymbol{x}_i$	$\alpha_i - 1$ occurrences of category i[note 1]	$\mathrm{DirMult}(\tilde{\boldsymbol{x}} \mid \boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size, N	M (number of target members)	Beta-binomial[4]	$n = N, \alpha, \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	α, β	$\alpha + n,\ \beta + \sum_{i=1}^{n} x_i - n$	$\alpha - 1$ experiments, $\beta - 1$ total failures[note 1]	

please check: https://en.wikipedia.org/wiki/Conjugate_prior (source)

Beta distribution

- conjugate prior for the Bernoulli, binomial, negative binomial and geometric distributions
- $Beta(\theta|a,b) \sim \theta^{a-1}(1-\theta)^{b-1}$



Beta distribution

- required $a, b > 0$
- if $a = b = 1$, we get the uniform distirbution
- if a and b are both less than 1, we get a bimodal distribution with "spikes" at 0 and 1
- if a and b are both greater than 1, the distribution is unimodal

Distribution properties

mean=$\frac{a}{a+b}$, mode=$\frac{a-1}{a+b-2}$, var=$\frac{ab}{(a+b)^2(a+b+1)}$

Beta prior

$Beta(\theta|a,b) \sim \theta^{a-1}(1-\theta)^{b-1}$

- prior parameters a and b are called **hyper-parameters**
- set a and b to encode your prior belief

Example

- to encode our beliefs that θ has mean 0.7 and standard deviation 0.2, we set a =2.975 and b =1.275
- to encode our beliefs that θ has mean 0.15 and that we think it lives in the interval (0.05, 0.30), we find a =4.5 and b =25.5

Posterior

Multiply the likelihood by the beta prior:

$$p(\theta|\mathcal{D}) \propto Bin(N_1|\theta, N_0 + N_1)Beta(\theta|a, b) \propto Beta(\theta|N_1 + a, N_0 + b)$$

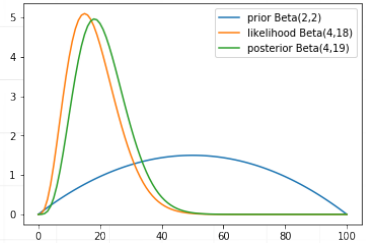


Figure: Beta(2,2) prior updated with Binomial likelihood with sufficient statistics $N_1 = 3$,$N_0 = 17$ yealds Beta(5,19)

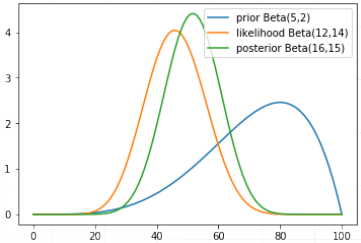


Figure: Beta(5,2) prior updated with Binomial likelihood with sufficient statistics $N_1 = 11$,$N_0 = 13$ yealds Beta(16,15)

Table of Contents

Frequentionist Linear Regression

$$y = \beta^T X + \epsilon$$

where:

- x represents predictor variables
- y is the response variable (also called the dependent variable)
- β are the weights (known as the model parameters)
- ϵ is an error term representing random sampling noise

GOAL: learning the coefficients β that best explain the data, i.e. minimize the residual sum of squares (RSS)

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - \hat{y})^2 = \sum_{i=1}^{N} (y_i - \beta^T x_i)^2$$

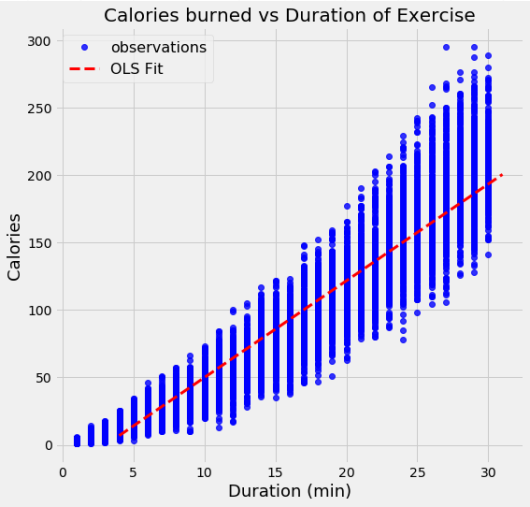
SOLUTION: Maximum Likelihood Estimate of β

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Frequentionist Linear Regression

Having $\hat{\beta}$

$$\hat{y} = \hat{\beta}^T X$$



Bayesian Linear Regression

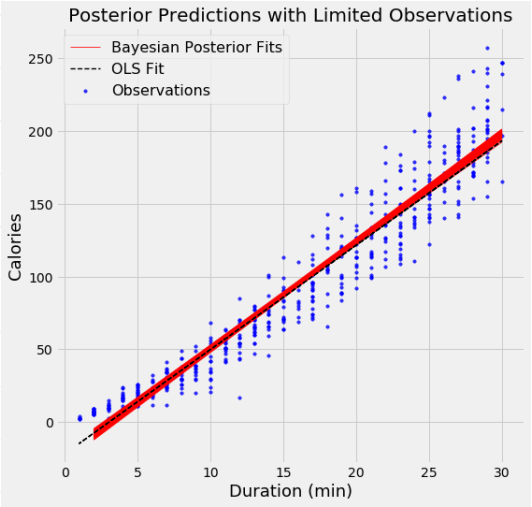
Here y is response sampled from Gaussian distrubution

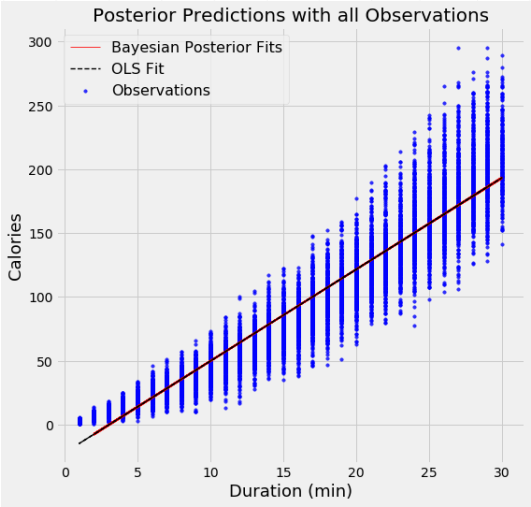
$$y \sim \mathcal{N}(\beta^T X, \sigma^2 I)$$

The aim is:
not to find the single "best" value of the model parameters
but to **determine the posterior distribution** for the model parameters
Posterior probability of the model parameters is conditional upon the training inputs and outputs:

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

Bayesian Linear Regression





Bayesian Linear Regression

Posterior Probability Density of Calories Burned from Bayesian Model

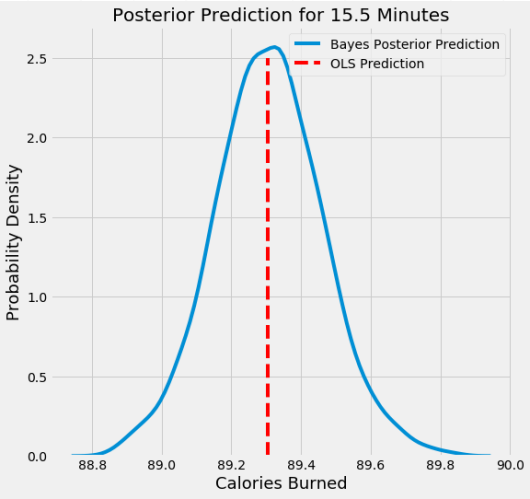
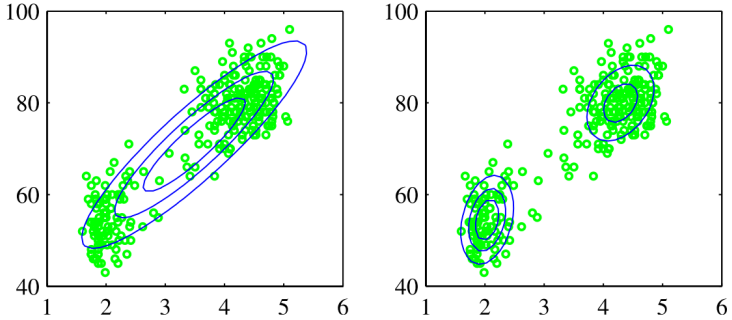


Table of Contents

Limmitations of Gaussian distribution

Gaussian distribution:

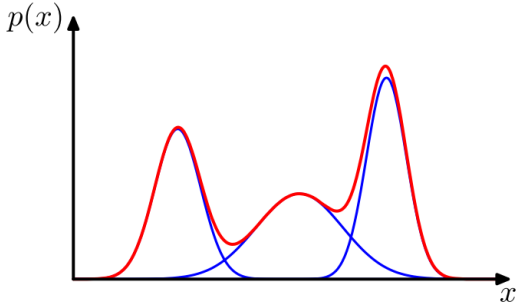
- has some important analytical properties
- but suffers from significant limitations when it comes to modelling real data sets.



- a linear superposition of two Gaussians gives a better characterization than single one

Gaussian Mixture Model

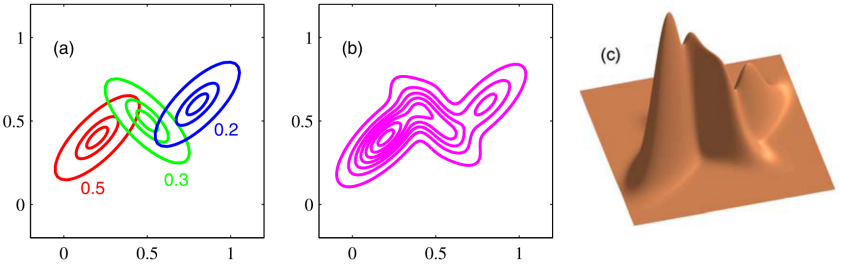
- using a sufficient number of Gaussians almost any continuous density can be approximated to arbitrary accuracy
- needed adjusting their means and covariances as well as the coefficients in the linear combination



$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

Mixture Gaussian

- each Gaussian density $\mathcal{N}(x|\mu_k, \Sigma_k)$ is called *component* of the mixture
- each has own mean μ_k and covariance Σ_k
- α_k are mixing coefficients
- $\Sigma_{k=1}^{K} \alpha_k = 1$
- $0 \leq \alpha_k \leq 1$



Mixture Gaussian

Mixing coefficients satisfy the requirements to be probabilities

Marginal density

$$p(x) = \sum_{k=1}^{K} p(k)p(x|k)$$

- $\alpha_k = p(k)$ viewd as prior probability of picking the kth component
- density $\mathcal{N}(x|\mu_k, \Sigma_k) = p(x|k)$ viewd as probability of x conditioned on k

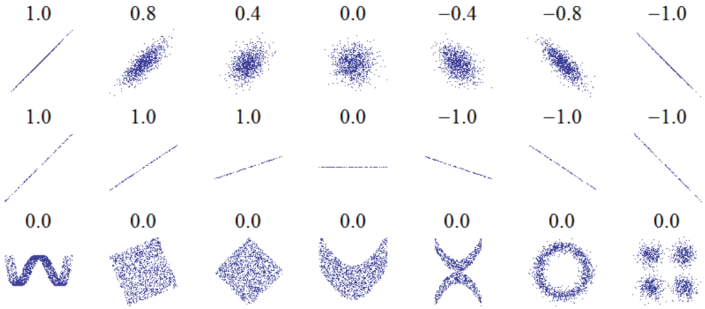
Table of Contents

Multivariate Gaussian (intro)

RECAP: Covariance

The covariance between two random vaiables X and Y measures the degree to which X and Y are (linearly) related.

$$cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$



Multivariate Gaussian (intro)

If x is a d-dimensional random vector, its **covariance matrix** is symmetric and positive definite:

$$cov[x] = \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T]$$
$$= \begin{pmatrix} var[X_1] & cov[X_1, X_2] & \cdots & cov[X_1, X_d] \\ cov[X_2, X_1] & var[x_2] & \cdots & cov[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ cov[X_d, X_1] & cov[X_d, X_2] & \cdots & var[X_d] \end{pmatrix}$$

- covariances can be between 0 and infinity
- more convenient to work with a normalized measure (a finite upper bound) ->

(Pearson) correlation coefficient

$$corr[X, Y] = \frac{cov[X, Y]}{\sqrt{var[X]var[Y]}}$$

Multivariate Gaussian

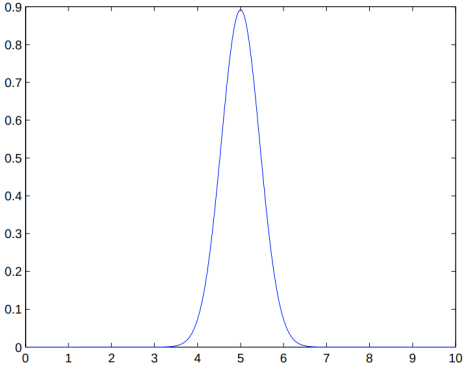
- most widely used joint probability density function for continuous random variables
- probability density function:

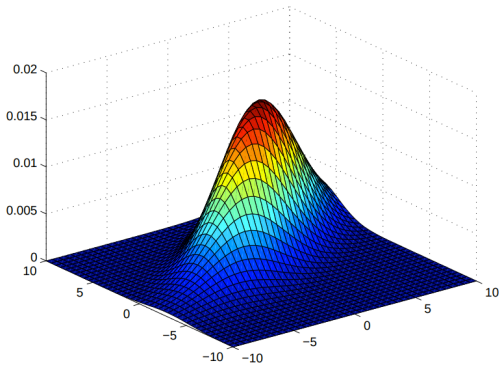
$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

  where $|\cdot|$ is the determinant, $\Sigma = cov[x]$ is $d \times d$ covariance matrix
- sometimes we can consider **precision matrix** (concentration matrix)

$$\Lambda = \Sigma^{-1}$$

Multivariate Gaussian





Homework 1

- Show that $corr[X, Y] = 1 \iff Y = aX + b$ for some a and b
- Calculate $cov[X, Y]$ and $corr[X, Y]$ if X and Y are independent ($P(X, Y) = P(X)P(Y)$)
- Does uncorrelated random variables imply their independence?
- What is *mutual information*? Is it better than correlation?