Machine Learning

Distributions

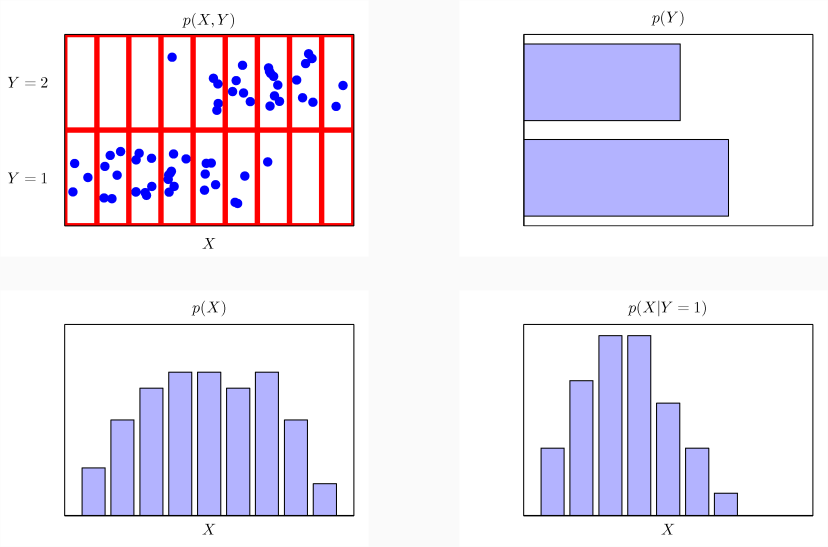
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 7, 2018

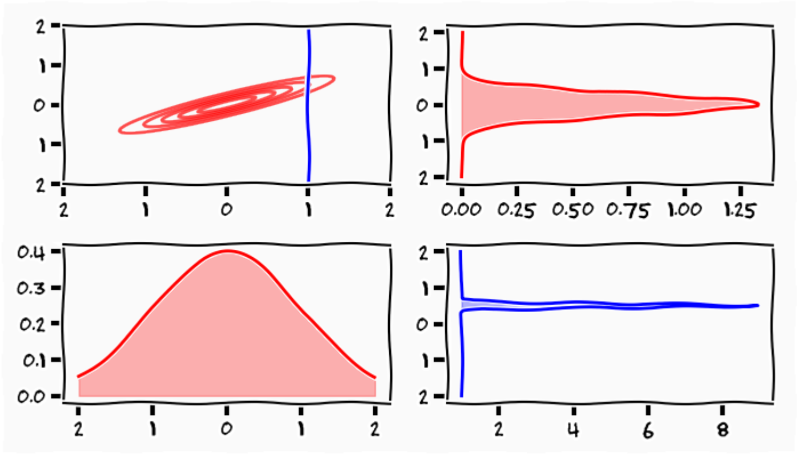
http://www.carlhenrik.com

Introduction

Basic Probabilities



Basic Probabilities



The Rules of Probability

Sum Rule

$$p(X) = \sum_{Y} p(X, Y)$$

Product Rule

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

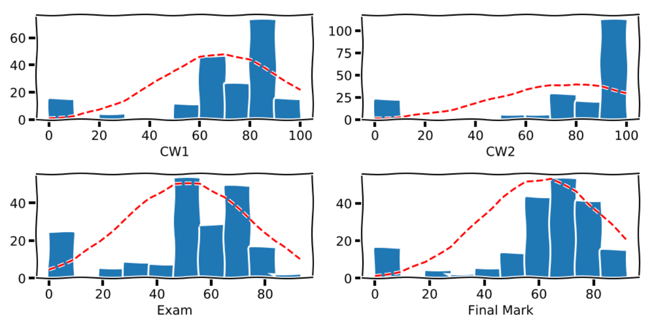
$\Rightarrow$ **Bayes Rule**

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

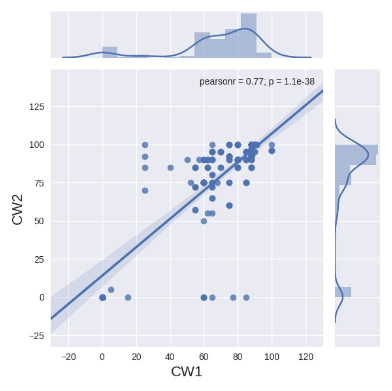
$$p(\mathrm{CW1}, \mathrm{CW2}, \mathrm{Exam})$$

- We have three courseworks in this unit
- *hopefully* they all relate to how good you are at Machine learning
- We are all very interested in asking questions about it
- Can't plot joint so lets look at some marginals from last year

Marginals



Marginal



$$p(\mathrm{CW1}, \mathrm{CW2}) = \sum_{x=1}^{100} p(\mathrm{CW1}, \mathrm{CW2}, \mathrm{Exam} = x) = \sum_{x=1}^{100} p(\mathrm{CW1}, \mathrm{CW2} | \mathrm{Exam} = x) p(\mathrm{Exam} = x)$$

Questions

Exam

$$p(\text{Exam} = 100 | \text{CW1} = 20, \text{CW2} = 30)$$

- What is the probability of me getting Exam=100 if CW1=20 and CW2=30
- As you will get a result on the exam the probability for all results sums to 1

$$\sum_{x=0}^{x=100} p(\text{Exam} = x | \text{CW1} = 20, \text{CW2} = 30) = 1.0$$

Questions

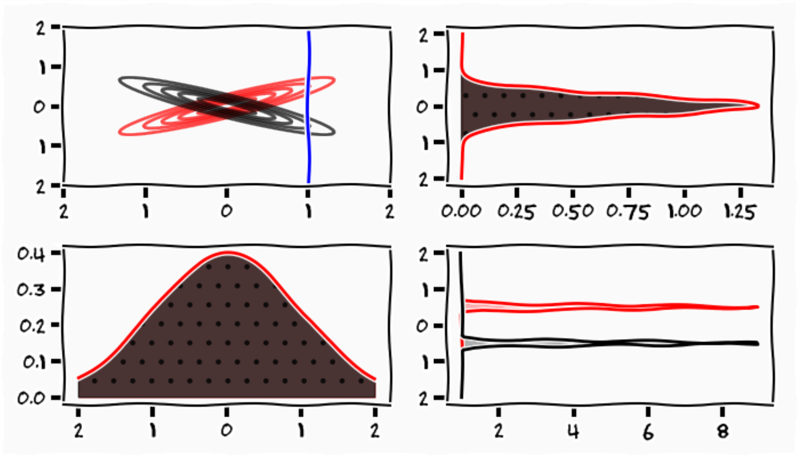
- Remember that each conditional is a probability
- However rare it is that you get 100% on both courseworks the conditional probability over all possibe exam results will sum to one

$$\sum_{x=1}^{x=100} p(\text{Exam} = x | \text{CW1} = 100, \text{CW2} = 100) = 1.0$$

- What shows that it is rare is that the probability for getting

$$\sum_{x=1}^{x=100} p(\text{Exam} = x, \text{CW1} = 100, \text{CW2} = 100)$$
$$= p(\text{CW1} = 100, \text{CW2} = 100) \leq 1.0$$

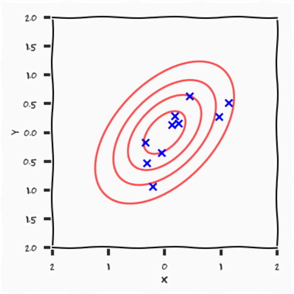
Dangers of Marginals



Dangers of Marginals

- Good looking people are paid more

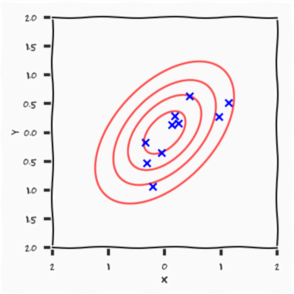
Learning with Distributions



- Our goal is to understand realisations of a system

[1] https://en.wikipedia.org/wiki/All_models_are_wrong

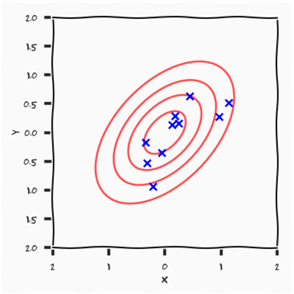
Machine Learning



- Our goal is to understand realisations of a system
- If we can, then we can "equate" our model with the system

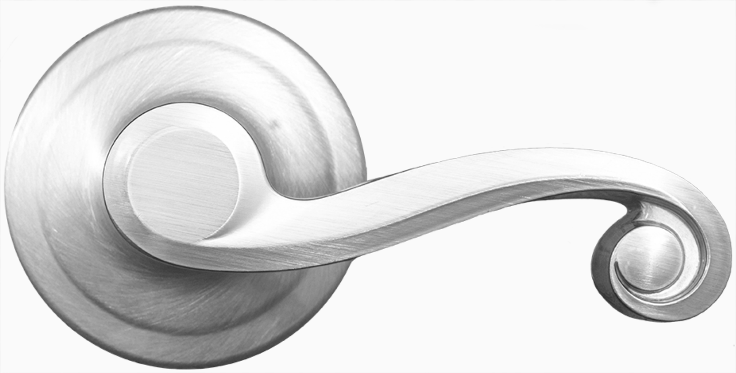
[1] https://en.wikipedia.org/wiki/All_models_are_wrong

Machine Learning



- Our goal is to understand realisations of a system
- If we can, then we can "equate" our model with the system
- Importantly not as truth, but as a useful hypothesis related to our assumptions[1]

[1] https://en.wikipedia.org/wiki/All_models_are_wrong



Bayes Rule

$$\underbrace{p(\theta|Y)}_{\text{posterior}} = \underbrace{P(Y|\theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta)}_{\text{prior}} \cdot \frac{1}{\underbrace{p(Y)}_{\text{evidence}}}$$

Likelihood how likely is the data to come from the model specific model indexed by θ

Bayes Rule

$$\underbrace{p(\theta|Y)}_{\text{posterior}} = \underbrace{P(Y|\theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta)}_{\text{prior}} \cdot \frac{1}{\underbrace{p(Y)}_{\text{evidence}}}$$

Likelihood how likely is the data to come from the model specific model indexed by θ

Prior what do I believe the specific model to be, i.e. how likely to I believe different θ to be

Bayes Rule

$$\underbrace{p(\theta|Y)}_{\text{posterior}} = \underbrace{P(Y|\theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta)}_{\text{prior}} \cdot \frac{1}{\underbrace{p(Y)}_{\text{evidence}}}$$

Likelihood how likely is the data to come from the model specific model indexed by θ

Prior what do I believe the specific model to be, i.e. how likely to I believe different θ to be

Evidence how likely do I think the data to be under all models weighted by how likely I think the specific models are

Bayes Rule

$$\underbrace{p(\theta|Y)}_{\text{posterior}} = \underbrace{P(Y|\theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta)}_{\text{prior}} \cdot \frac{1}{\underbrace{p(Y)}_{\text{evidence}}}$$

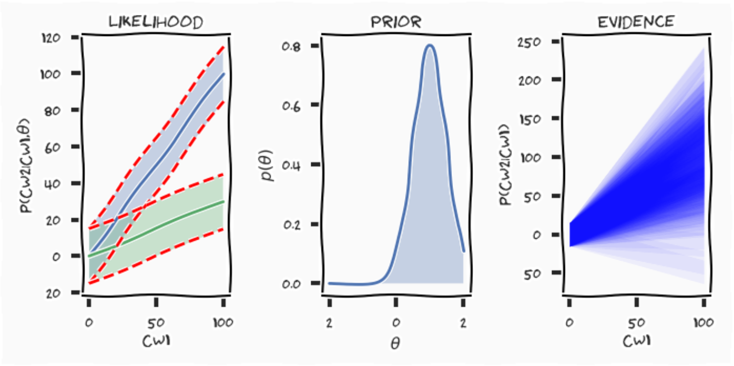
Likelihood how likely is the data to come from the model specific model indexed by θ

Prior what do I believe the specific model to be, i.e. how likely to I believe different θ to be

Evidence how likely do I think the data to be under all models weighted by how likely I think the specific models are

Posterior which distributions of models do I believe have generated this data

Machine Learning



$$\text{CW2} = \theta \cdot \text{CW1} \pm 15\%$$

$$\theta \sim \mathcal{N}(1.0, 0.5)$$

Discrete Distributions

Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

- Due to binary outcome

$$p(x = 0|\mu) = 1 - \mu$$

Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

- Due to binary outcome

$$p(x = 0|\mu) = 1 - \mu$$

- Distribution

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}$$

Data

- We want to figure out what μ is for a specific coin
- Toss the coin N times, $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- What happens if we blindly trust this one experiment?

Maximum Likelihood

$$\mu_{ML} = \text{argmax}_{\mu} p(\mathcal{D}|\mu) = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- if we get 3 heads in a row, we believe it will always be heads
- we need to include an assumption as a prior over μ

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})}$$

- Also gives us an uncertainty related to our knowledge

Posterior

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})}.$$

- if we can specify a prior $p(\mu)$ we can reach the posterior belief

Posterior

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})}.$$

- if we can specify a prior $p(\mu)$ we can reach the posterior belief
- what do we know about coins?

Posterior

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})}.$$

- if we can specify a prior $p(\mu)$ we can reach the posterior belief
- what do we know about coins?
- how do I make that knowledge mathematicall explicit?

Conjugate Prior

- If we have a prior belief μ we want the posterior belief to have the same functional form

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

Conjugate Prior

- If we have a prior belief μ we want the posterior belief to have the same functional form

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

- Likelihood

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

Conjugate Prior

- If we have a prior belief μ we want the posterior belief to have the same functional form

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

- Likelihood

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- Conjugate prior

$$p(\mu|\theta) = f_1(\theta)\mu^{f_2(\theta)}(1-\mu)^{f_3(\theta)}$$

$$\int_0^1 p(\mu|\theta)\mathrm{d}\mu = 1$$

Conjugate Prior

- If we have a prior belief μ we want the posterior belief to have the same functional form

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

- Likelihood

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

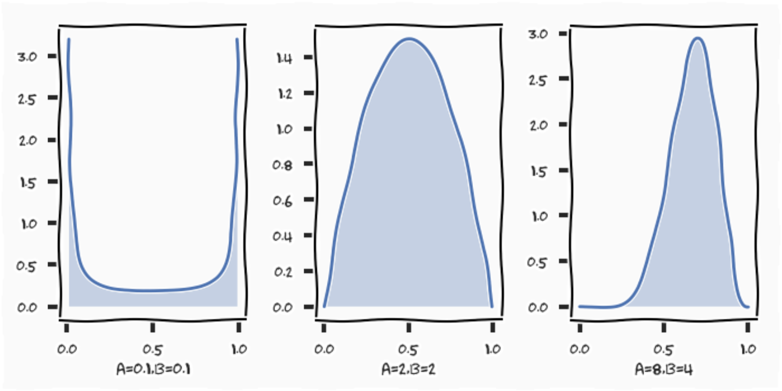
- Conjugate prior

$$p(\mu|\theta) = f_1(\theta)\mu^{f_2(\theta)}(1-\mu)^{f_3(\theta)}$$

$$\int_0^1 p(\mu|\theta)\mathrm{d}\mu = 1$$

- *Does this make philosophical sense?*

Beta Distribution



$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

Assumption 1: Independence

$$p(\mathbf{x}|\mu) = \prod_{i=1}^{N} \mathrm{Bern}(x|\mu) = \prod_{i=1}^{N} \mu^{x}(1-\mu)^{1-x}.$$

Lets assume that each toss of the coin is independent

Posterior

Churn the handle

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mu)p(\mu)}{\underbrace{\int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu}_{\text{This is hard}}}$$

Posterior

Churn the handle

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mu)p(\mu)}{\underbrace{\int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu}_{\text{This is hard}}}$$

Conjugacy

- We know the functional form of the posterior

Posterior

Churn the handle

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mu)p(\mu)}{\underbrace{\int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu}_{\text{This is hard}}}$$

Conjugacy

- We know the functional form of the posterior
- We know that the posterior is proportional to the likelihood times the prior

Posterior

Churn the handle

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mu)p(\mu)}{\underbrace{\int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu}_{\text{This is hard}}}$$

Conjugacy

- We know the functional form of the posterior
- We know that the posterior is proportional to the likelihood times the prior
- *Use these facts to avoid the integral*

Posterior

$$p(\mu|\mathcal{D}) \propto p(\mathcal{D}|\mu)p(\mu)$$

Posterior

$$
\begin{aligned}
p(\mu|\mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu) \\
&= \prod_{i=1}^{N} \mathrm{Bern}(x|\mu)\mathrm{Beta}(\mu|a, b)
\end{aligned}
$$

Posterior

$$
\begin{aligned}
p(\mu|\mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu) \\
&= \prod_{i=1}^{N} \mathrm{Bern}(x|\mu)\mathrm{Beta}(\mu|a,b) \\
&= \prod_{i=1}^{N} \mu^{x}(1-\mu)^{1-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}
\end{aligned}
$$

Posterior

$$
\begin{aligned}
p(\mu|\mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu) \\
&= \prod_{i=1}^{N} \mathrm{Bern}(x|\mu)\mathrm{Beta}(\mu|a,b) \\
&= \prod_{i=1}^{N} \mu^{x}(1-\mu)^{1-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \mu^{\sum_i x_i}(1-\mu)^{\sum_i (1-x_i)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}
\end{aligned}
$$

Posterior

$$
\begin{aligned}
p(\mu|\mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu) \\
&= \prod_{i=1}^{N} \mathrm{Bern}(x|\mu)\mathrm{Beta}(\mu|a,b) \\
&= \prod_{i=1}^{N} \mu^{x}(1-\mu)^{1-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \mu^{\sum_i x_i}(1-\mu)^{\sum_i (1-x_i)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{\sum_i x_i + a}(1-\mu)^{\sum_i (1-x_i)+b-1}.
\end{aligned}
$$

Posterior

- Because we know the form of the posterior, we can *identify* its parameters

$$\text{Beta}(\mu|a_n, b_n) \propto \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{\underbrace{\sum_i x_i + a}_{a_n}} (1-\mu)^{\underbrace{\sum_i (1-x_i) + b}_{b_n} -1}$$

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

Posterior

- Because we know the form of the posterior, we can *identify* its parameters

$$\text{Beta}(\mu|a_n, b_n) \propto \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{\underbrace{\sum_i x_i + a}_{a_n}} (1-\mu)^{\underbrace{\sum_i (1-x_i) + b}_{b_n} -1}$$

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

- This leads to the following posterior

$$\text{Beta}(\mu|a_n, b_n) = \frac{\Gamma\left(\sum_i x_i + a + \sum_i (1-x_i) + b\right)}{\Gamma\left(\sum_i x_i + a\right)\Gamma\left(\sum_i (1-x_i) + b\right)} \mu^{\sum_i x_i + a}(1-\mu)^{\sum_i (1-x_i) + b - 1}$$

Document

`Lectures/bernoullitrial.pdf`

- Have a look at this document
- Implement the code (its listed) see if you get the intuition

Multinomial

- If we have a variable that can take K different states

$$\mathbf{x} = [0, 0, 1, 0, 0, 0]^{\mathrm{T}}$$

Multinomial

- If we have a variable that can take K different states

$$\mathbf{x} = [0, 0, 1, 0, 0, 0]^{\mathrm{T}}$$

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]^{\mathrm{T}}, \sum_k \mu_k = 1$$

Multinomial

- If we have a variable that can take K different states

$$\mathbf{x} = [0, 0, 1, 0, 0, 0]^{\mathrm{T}}$$

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]^{\mathrm{T}}, \sum_k \mu_k = 1$$

- Likelihood

$$p(\mathbf{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}}$$

Dirichlet

- Multinomial

$$p(\mathsf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

Dirichlet

- Multinomial

$$p(\mathsf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- Conjugate prior

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

Dirichlet

- Multinomial

$$p(\mathsf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- Conjugate prior

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

- Dirichlet Distribution

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdot \ldots \cdot \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

Dirichlet

- Posterior

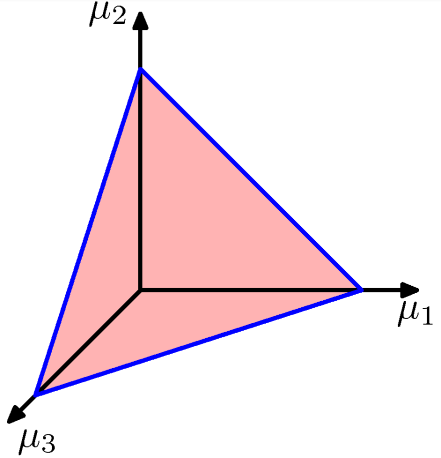
$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k+1}$$

$$m_k = \sum_n x_{nk}$$

- Normalised Form

$$p(|\mathcal{D},\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdot\ldots\cdot\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k+1}$$

Dirichlet Prior



Spans the plane $\mu_1 + \mu_2 + \mu_3 = 1$

Hyperparameters

$$p(\mu|\mathcal{D}, \boldsymbol{\alpha}) = \frac{p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha})}{p(\mathcal{D}|\boldsymbol{\alpha})}$$

- all these priors have parameters, where do they come from?

Hyperparameters

$$p(\mu|\mathcal{D}, \boldsymbol{\alpha}) = \frac{p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha})}{p(\mathcal{D}|\boldsymbol{\alpha})}$$

- all these priors have parameters, where do they come from?
- either we know them

Hyperparameters

$$p(\mu|\mathcal{D}, \boldsymbol{\alpha}) = \frac{p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha})}{p(\mathcal{D}|\boldsymbol{\alpha})}$$

- all these priors have parameters, where do they come from?
- either we know them
- if we don't then place a prior over the priors parameters and go again

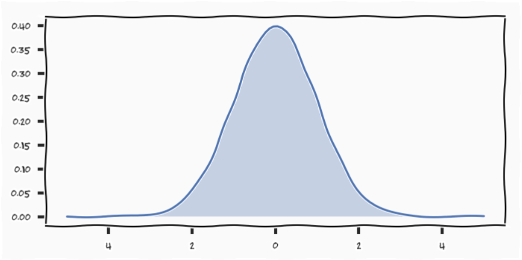
Hyperparameters

$$p(\mu|\mathcal{D}, \boldsymbol{\alpha}) = \frac{p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha})}{p(\mathcal{D}|\boldsymbol{\alpha})}$$

- all these priors have parameters, where do they come from?
- either we know them
- if we don't then place a prior over the priors parameters and go again
- the idea is to build up a hierarchy until you can input your knowledge/assumptions

Continous Distributions

Gaussian Distribution



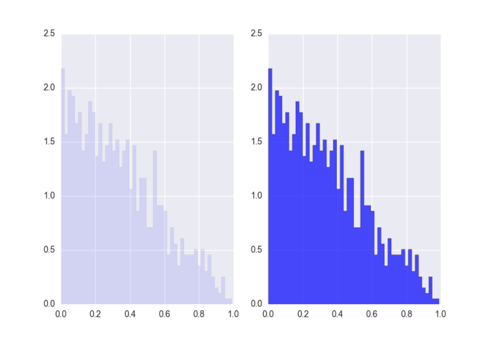
$$p(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

Central Limit Theorem[2]

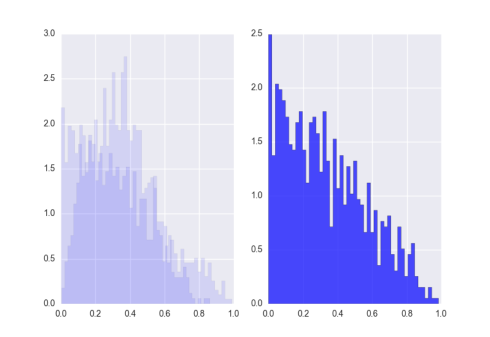
The central limit theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.

[2] https://www.youtube.com/watch?v=wadzsURQFT4

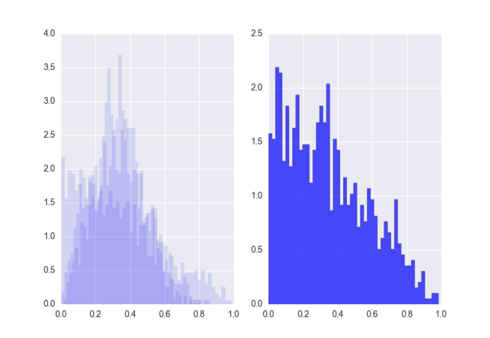
Central Limit Theorem



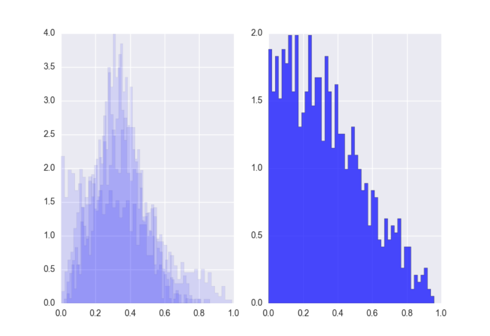
Central Limit Theorem



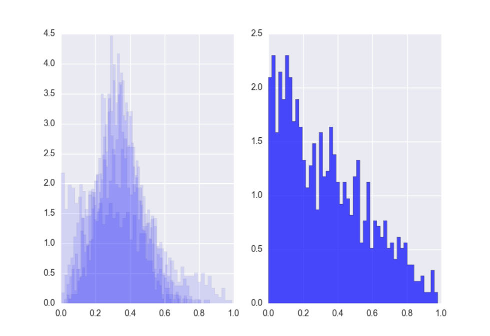
Central Limit Theorem



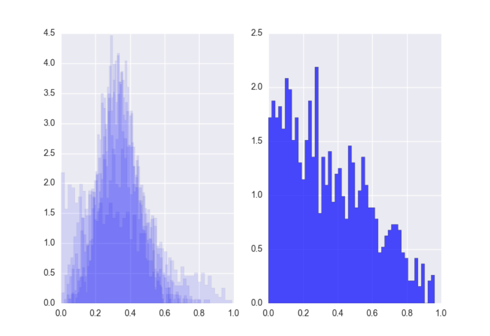
Central Limit Theorem



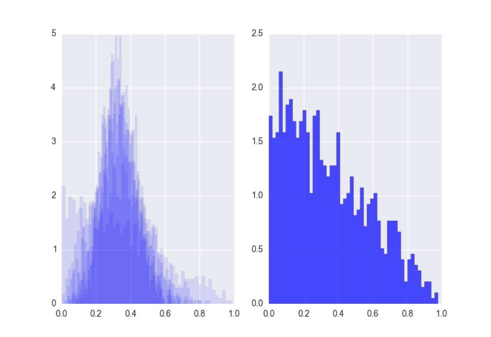
Central Limit Theorem



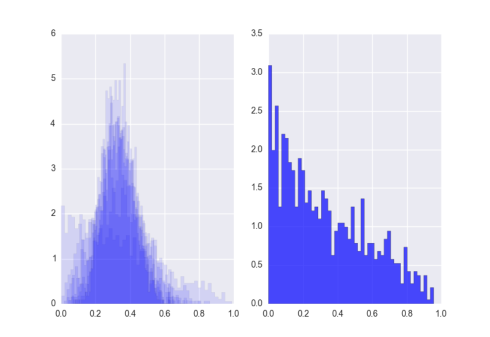
Central Limit Theorem



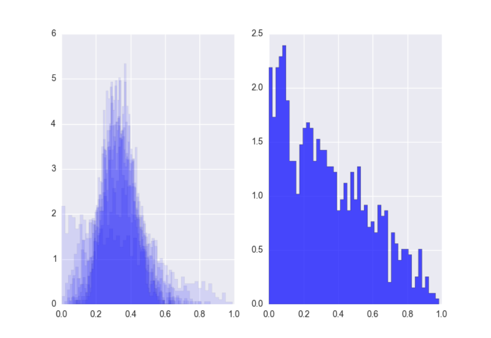
Central Limit Theorem



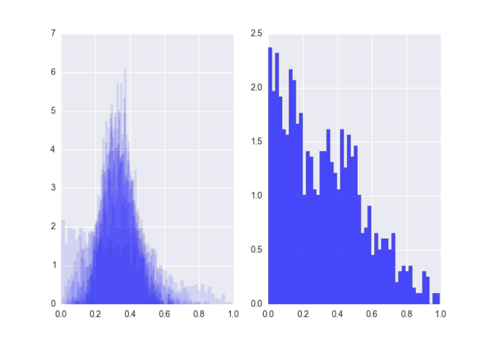
Central Limit Theorem



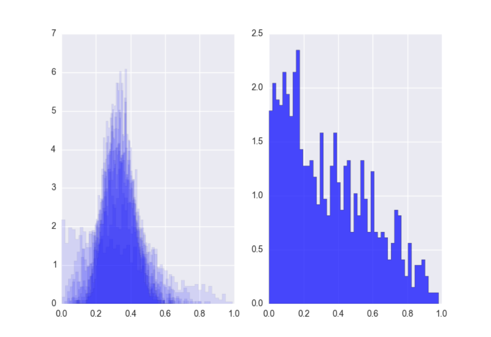
Central Limit Theorem



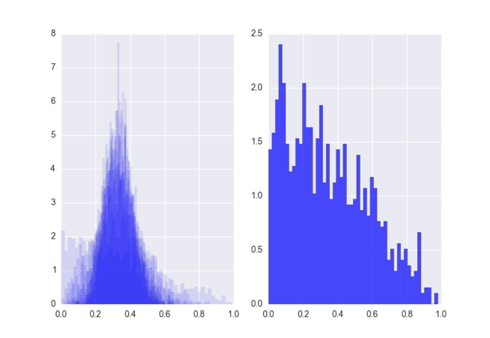
Central Limit Theorem



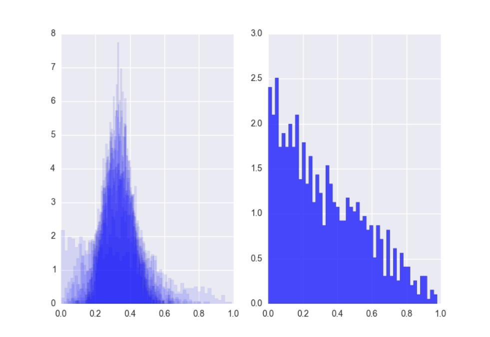
Central Limit Theorem



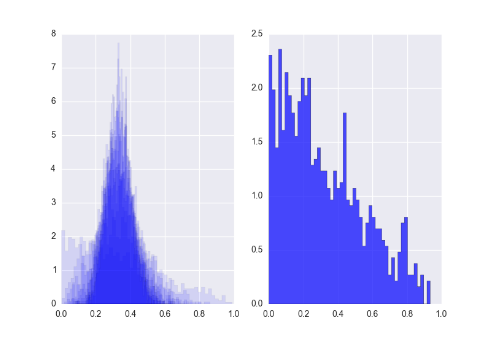
Central Limit Theorem



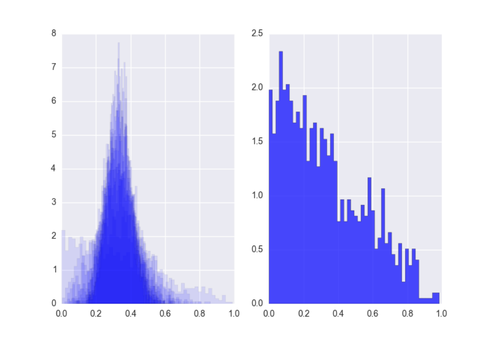
Central Limit Theorem



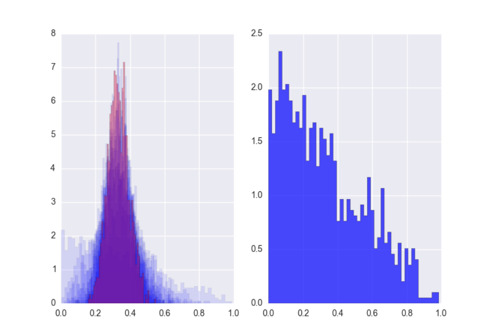
Central Limit Theorem



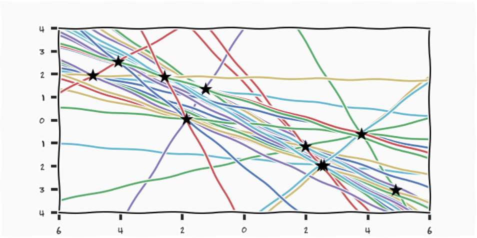
Central Limit Theorem



Central Limit Theorem



Cerces



The search for Cerces

Gauss made the assumption that Piazzi's measurment errors where *independent* draws from a *unknown* distribution that was *fixed*. This we often know as `i.i.d` *Independent and Identically Distributed*

Conjugate Prior[3]

- Gaussians are self-conjugate
  - Gaussian likelihood + Gaussian Prior $\rightarrow$ Gaussian Posterior
- Gaussian distribution
  - Conjugate prior for $\boldsymbol{\mu}$ is Gaussian
  - Conjugate prior for $\boldsymbol{\Sigma}$ is Inverse-Wishard

[3] https://en.wikipedia.org/wiki/Conjugate_prior

Gaussian Distribution

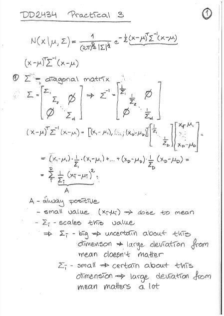
$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

Posterior $p(x_1|x_2) \propto p(x_2|x_1)p(x_1)$

Marginal $p(x_1) = \int p(x_1, x_2)\mathrm{d}x_2$

Conditional $p(x_1|x_2) = \frac{p(x_1,x_2)}{p(x_2)}$

Gaussian Identities



DD2434 Practical 3 ①

$$N(x|\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}\underbrace{(x-\mu)^T\Sigma^{-1}(x-\mu)}}$$

$$(x-\mu)^T\Sigma^{-1}(x-\mu)$$

① Σ = diagonal matrix

$$\Sigma=\begin{bmatrix}\Sigma_1 & & \emptyset\\ & \Sigma_2 & \\ \emptyset & \ddots & \Sigma_d\end{bmatrix}\Rightarrow\Sigma^{-1}=\begin{bmatrix}\frac{1}{\Sigma_1} & & \emptyset\\ & \frac{1}{\Sigma_2} & \\ \emptyset & \ddots & \frac{1}{\Sigma_d}\end{bmatrix}$$

$$(x-\mu)^T\Sigma^{-1}(x-\mu)=[(x_1-\mu_1),\ldots,(x_D-\mu_D)]\begin{bmatrix}\frac{1}{\Sigma_1} & \\ & \ddots \\ & & \frac{1}{\Sigma_D}\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ \vdots\\ x_D-\mu_D\end{bmatrix}=$$

$$=(x_1-\mu_1)\cdot\frac{1}{\Sigma_1}\cdot(x_1-\mu_1)+\ldots+(x_D-\mu_D)\cdot\frac{1}{\Sigma_D}(x_D-\mu_D)=$$

$$=\sum_i^D\underbrace{\frac{1}{\Sigma_i}(x_i-\mu_i)^2}_{A}$$

A - always positive
- small value $(x_i-\mu_i)\Rightarrow$ close to mean
- Σ_i - scales this value
  - $\Rightarrow\Sigma_i$ - big $\Rightarrow$ uncertain about this dimenson $\Rightarrow$ large deviation from mean doesn't matter
  - Σ_i - small $\Rightarrow$ certain about this dimension $\Rightarrow$ large deviation from mean matters a lot

Tuesday 17-until they kick us out

Exponential Family [1] Ch. 2.4

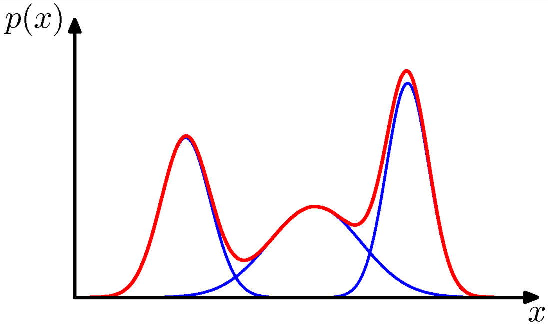
- Most distributions are parametetrised using exponentials
- Exponential family natural parametrisation

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})e^{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})}$$

- Conjugate prior

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)g(\boldsymbol{\chi})^{\nu}e^{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}}$$

Mixtures of Gaussians [1] Ch 2.3.9

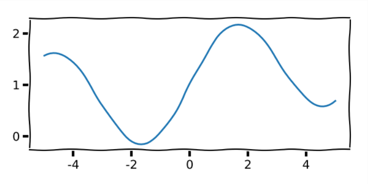


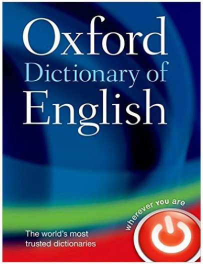
$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) \underbrace{p(\mathbf{x}|k)}_{\mathcal{N}(\mu_k, \Sigma_k)}$$

Stochastic Process

$$\infty$$

Stochastic Process





Kologrovs Existence Theorem

Defines what a distribution needs to forfill in order for a process to exist. Each finite instantiaton of the process is this distribution.

Summary

Summary

- Distributions allows us to make our assumptions explicit
- Conjugacy implies that the posterior and the prior is in the same family
- This finishes our introduction
- Now we have the tools that allows us to do Machine Learning

eof

Appendix

Posterior

$$p(x_1|x_2) \propto p(x_2|x_1)p(x_1)$$

1. Multiply right-hand side
2. Look at the exponents
3. Find the three terms, constant, mixed and quadratic
4. Complete the square to find the parameters

Marginal

$$p(x_1) = \int p(x_1, x_2)\mathrm{d}x_2 = \mathcal{N}(\mu_1, \Sigma_{11})$$

1. Write out the exponent of the joint distribution
2. Complete Square and collect terms with $x_1 - \mu_1$ (as we know the result)
3. Compute integral by knowing that densities always integrates to one

Conditional

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)}$$

1. Factorise the problem as $p(x_1, x_2) = p(x_1|x_2)p(x_2)$
2. We know the marginal and the joint
3. Use Schur complement to re-write the covariance matrix on block form

References

Christopher M. Bishop.
Pattern Recognition and Machine Learning (Information Science and Statistics).
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.