Machine Learning

Distributions

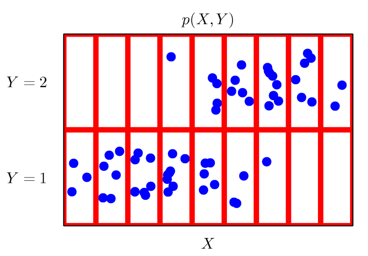
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

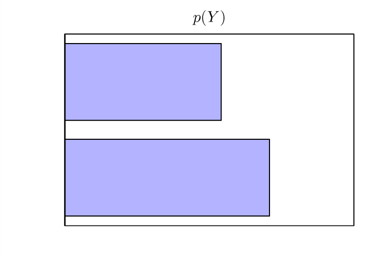
October 2, 2017

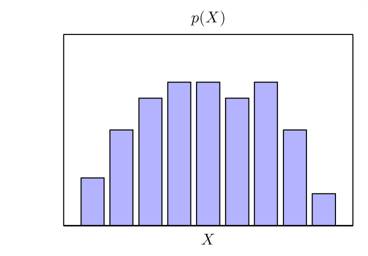
http://www.carlhenrik.com

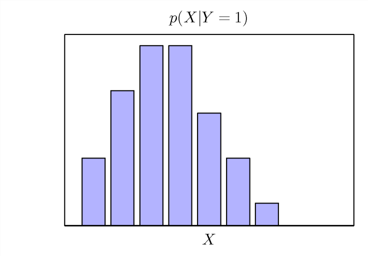
Introduction

Basic Probabilities

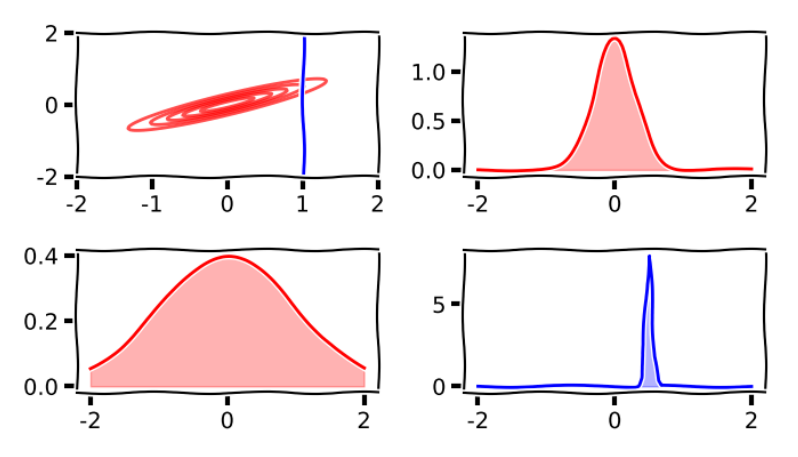








Basic Probabilities



The Rules of Probability

Sum Rule

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

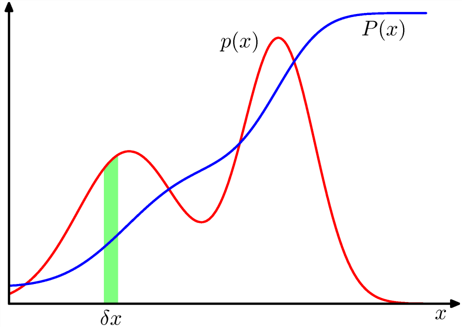
Product Rule

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

$\Rightarrow$ **Bayes Rule**

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

Probability Densities [1] ch 1.2.1



$$\lim_{\delta x \to 0} p(x \in (x, x + \delta x)) = \lim_{\delta x \to 0} \int_x^{x+\delta x} p(x)\mathrm{d}x = p(x) \cdot \delta x$$

$$p(x) \geq 0, \quad \int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$$

Machine Learning

- Our goal is to understand realisations of a system

Machine Learning

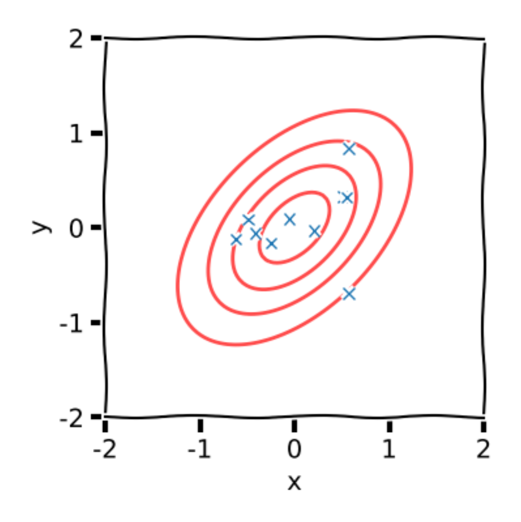
- Our goal is to understand realisations of a system
- If we can, then we can "equate" our model with the system

Machine Learning

- Our goal is to understand realisations of a system
- If we can, then we can "equate" our model with the system
- if you are observing a system and never get surprised, would you say that you understand the system?

Machine Learning

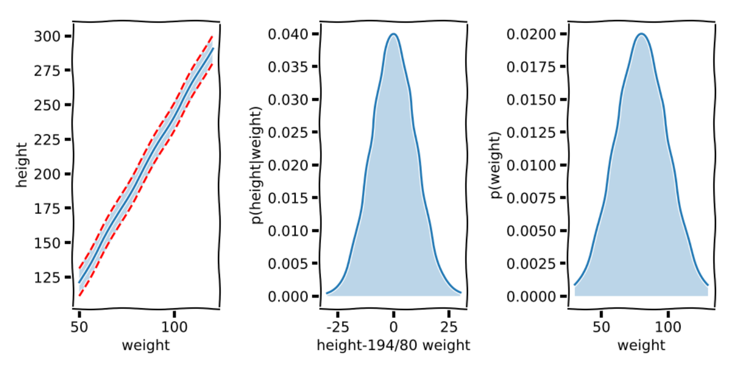
- Our goal is to understand realisations of a system
- If we can, then we can "equate" our model with the system
- if you are observing a system and never get surprised, would you say that you understand the system?
- *if you think of the probability as a measure of "suprisedness", if you have probability 0 and you see data you will be very surprised."*


2
1
0
-1
-2
y
-2
-1
0
1
2
x

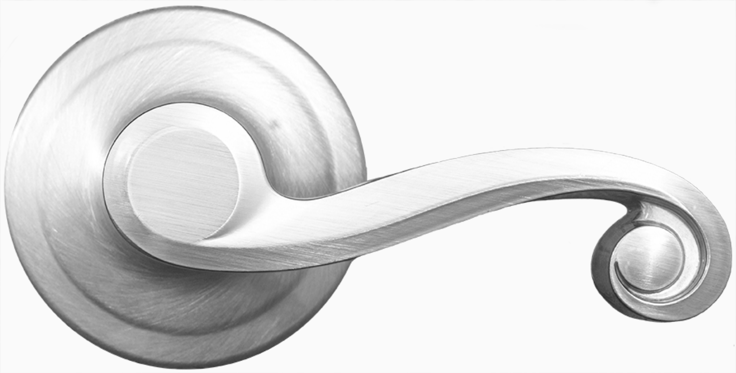
Bayes Rule

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

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



$$p(h|w) = \mathcal{N}(w \cdot \frac{194}{80}, 10^2)$$
$$p(w) = \mathcal{N}(80, 20^2)$$



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

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

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

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

- Binomial

$$\mathrm{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

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

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

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

- 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

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

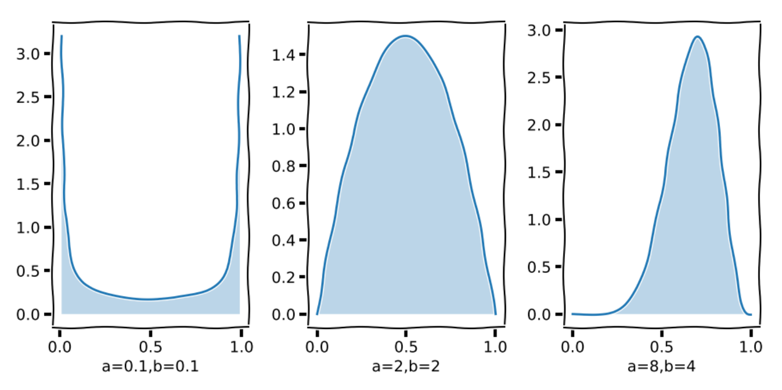
- Conjugate prior

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

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

- Does this make philosophical sense?

Beta Distribution



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

Binomial Posterior

$$
\begin{aligned}
p(\mu|m,a,b) &\propto p(\mathcal{D}|\mu)p(\mu) \\
&= \binom{N}{m}\mu^m(1-\mu)^{N-m}\cdot\frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&\propto \mu^{m+a-1}(1-\mu)^{N-m+b-1}
\end{aligned}
$$

- the parameters of the prior have a clear interpretation

  a number of extra observations of $x = 0$

  b number of extra observations of $x = 1$

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

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

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

- Likelihood

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

Dirichlet

- Multinomial

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

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

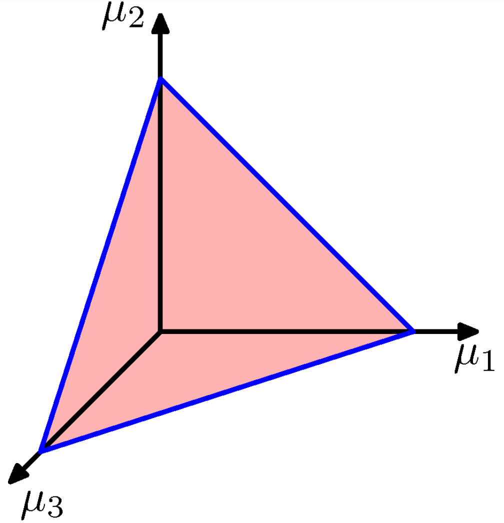
$$p(|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k + 1}$$

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

- 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



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

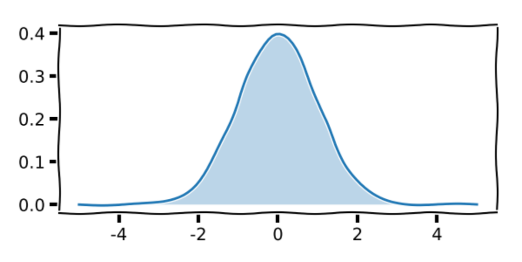
Hyperparameters

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

- 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

Continous Distributions

Gaussian Distribution



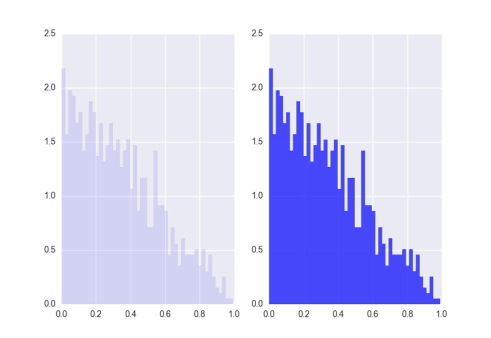
$$p(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

Central Limit Theorem[1]

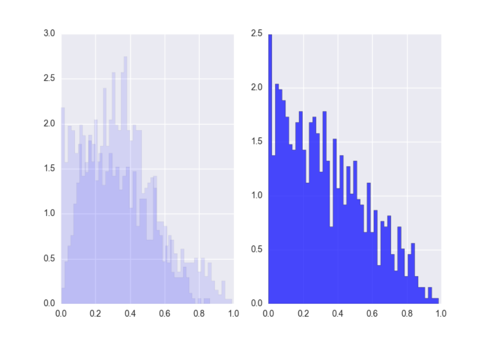
The central limit theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.

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

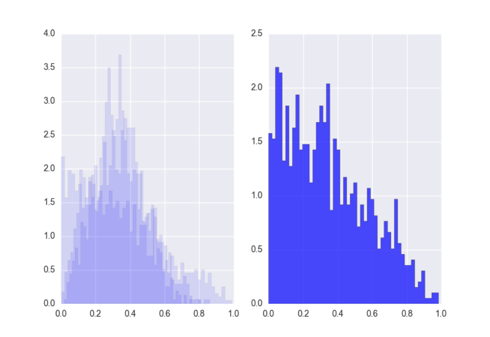
Central Limit Theorem



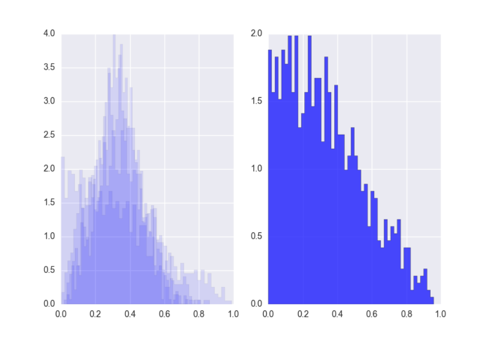
Central Limit Theorem



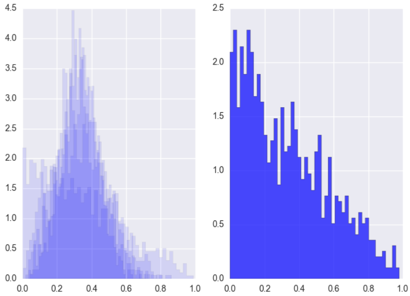
Central Limit Theorem



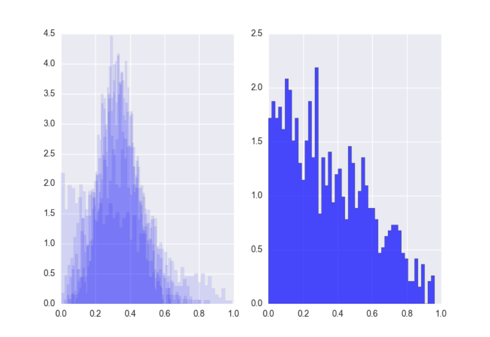
Central Limit Theorem



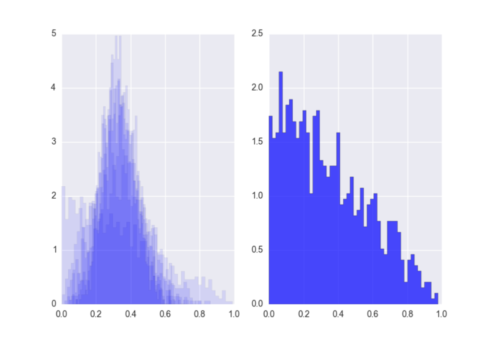
Central Limit Theorem



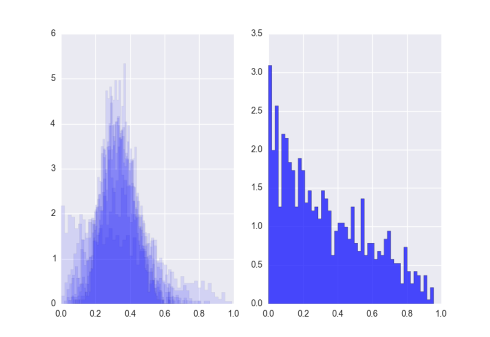
Central Limit Theorem



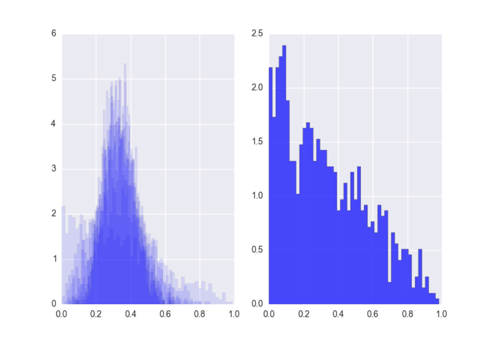
Central Limit Theorem



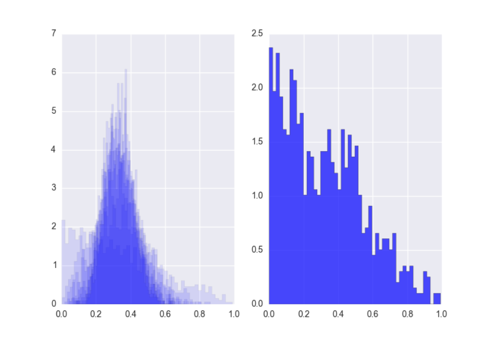
Central Limit Theorem



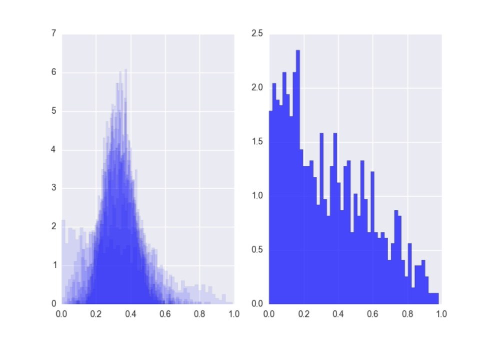
Central Limit Theorem



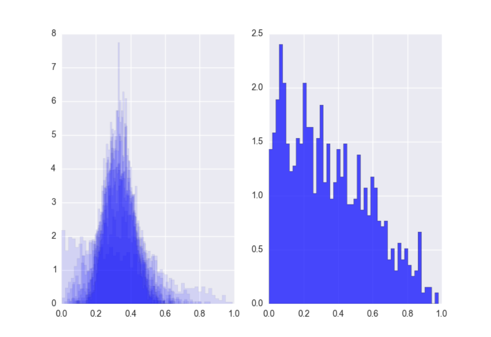
Central Limit Theorem



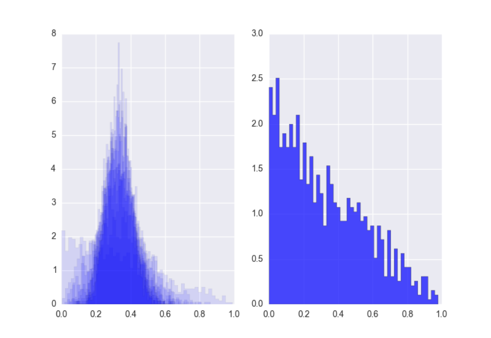
Central Limit Theorem



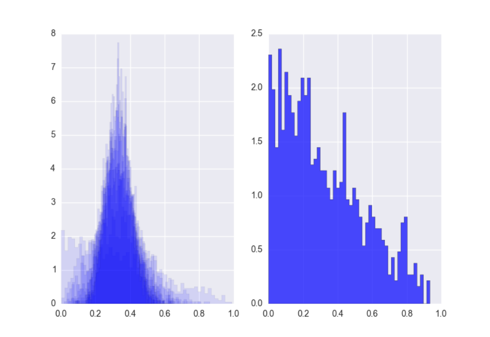
Central Limit Theorem



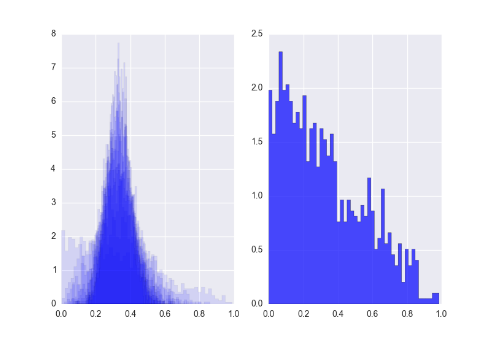
Central Limit Theorem



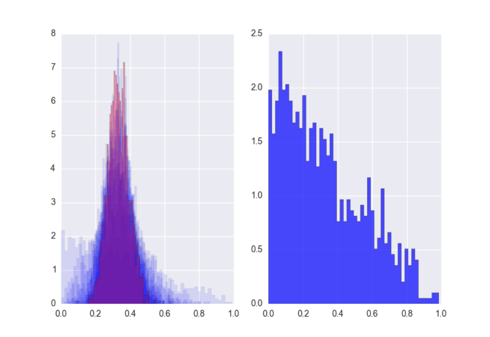
Central Limit Theorem



Central Limit Theorem



Central Limit Theorem



Central Limit Theorem Carl

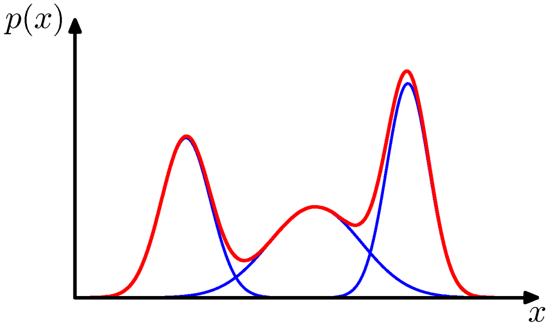
- If I do not know anything at all except that I think that there is some distribution something comes from
- A sensible idea (potentially) would be to think of an average
- As it turns out, the average of anything that I do not know (but are identically distributed) is Gaussian
- Therefore I think it kinda makes sense

Conjugate Prior[2]

- 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

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

Mixtures of Gaussians [1] Ch 2.3.9



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

- Exponential family natural parametrisation

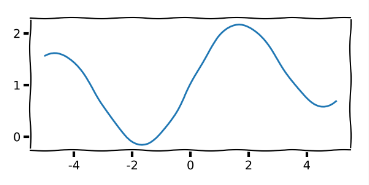
$$p(\mathsf{x}|\boldsymbol{\eta}) = h(\mathsf{x})g(\boldsymbol{\eta})e^{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathsf{x})}$$

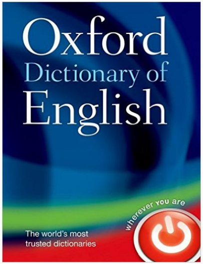
- Conjugate prior

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

Stochastic Process

$$\infty$$





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.

Gaussian Identities

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

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. Called completing the square and very very useful to have done

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

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

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

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

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

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

A - alway positive
- small value $(x_i-\mu_i) \rightarrow$ close to mean
- Σ_i - scales this value

$\Rightarrow$ Σ_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

Summary

Summary

- Distributions allows us to make our assumptions explicit
- Conjugacy implies that the posterior and the prior is in the same family
- Exponential family defines a *natural* parametrisation of distributions that we can work with
- Gaussian Identities!

eof

References

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