Probabilistic reasoning: *supervised learning*

Marco Piastra

Types of learning problems

Consider a number of observations (input data) made by an agent

$\{D_1, D_2, ..., D_n\}$

- ***Supervised learning***

  Learning form <u>*complete*</u> observations: together with the input objects $\{D_1, D_2, ..., D_n\}$, the agent knows a set of corresponding expected values $\{Y_1, Y_2, ..., Y_n\}$

  The objective is learning a *joint distribution* P

- ***Unsupervised learning***

  Learning form <u>*incomplete*</u> observations: from a set of incomplete observations $\{D_1, D_2, ..., D_n\}$ the agent wants to learn a complete ***model***

  The objective is learning a *joint distribution* P

- ***Reinforcement learning***

  The observations $\{D_1, D_2, ..., D_n\}$ are *states* o *situations*, at each state X_i the agent must perform an ***action*** a_i that produces a ***result*** r_i.

  The objective is defining a *function* $a_i = \pi(D_i)$ that describes a strategy that the agent will follow

  The strategy should be optimal, in the sense that it should maximize the expected value of a *function* $v(< r_1, r_2, ..., r_n >)$ of the sequence of *results*

Events and observations

- Events

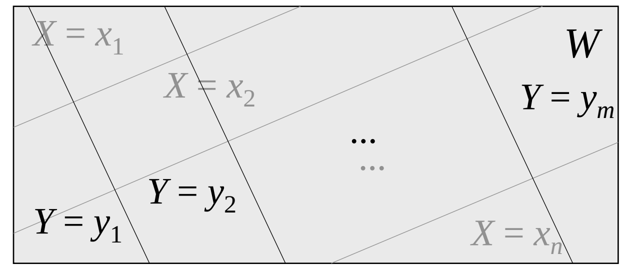
  An **event** is a subset of *possible worlds*

  An event **occurs** when the actual world is known to belong to the subset

- Multiple random variables

  A convenient way to define a σ-algebra of events

  In the discrete case, each combination of values of the random variables describes an *event*



- Observations (data)

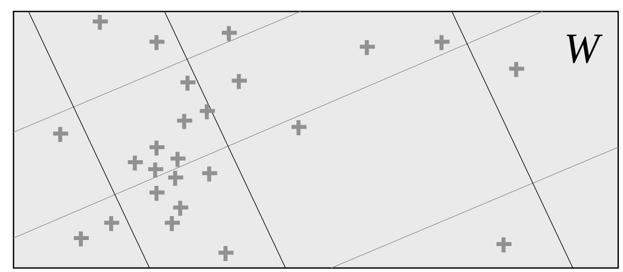
  Each *observation* is about one *possible world*

  In each possible world, all the values of random variables are determined

  Observations could be either *complete* or *partial*

  In the sense that not all the relevant values could be actually observed

  How do observations (complete or not) connect to *probability* P ?



Observations and Independence

About notation

Each observation could be the outcome of an experiment or a test

The outcome of a particular experiment will be represented by a set of *random variables*

For example, if the model adopts the random variable $\{X, Y\}$,
the n outcomes of the experiments are $D_1 = \{X_1, Y_1\}$, $D_2 = \{X_2, Y_2\}$, ..., $D_n = \{X_n, Y_n\}$

- *Independent observations, same probability distribution*

*Independent, Identically Distributed (**IID**) random variables*

Definition

A sequence o or set of random variables $\{X_1, X_2, \ldots, X_n\}$ is IID iff:

1) $<X_i \perp X_j>, i \neq j$ (independence)
2) $P(X_i) = P(X_i), i \neq j$ (same distribution)

The extension to sequences of subsets of random variable is immediate

CAUTION:

Being IID is not an obvious property of observations

Example: different measurements on different patients *may* be IID,
but different measurements over time on the same patient are <u>not</u> IID

Maximum Likelihood Estimation (MLE)

A probabilistic model $P(X)$, with parameters θ

θ is a vector of values that characterizes $P(X)$ *completely*: once θ is defined, $P(X)$ is also defined.

A set of IID observations $D = \{D_1, D_2, \dots , D_n\}$

- *Likelihood function*

A function, or a conditional probability, derived from the model $P(X)$

$$L(\theta | D) = P(D | \theta) = P(D_1, D_2, \ldots, D_n | \theta)$$

where $P(D | \theta)$ is the conditional probability that the parameter θ, considered as a random variables, could generate the observations D

When the observations $\{D_1, D_2, \dots , D_n\}$ are IID:

$$P(D | \theta) = P(D_1 | \theta) P(D_2 | \theta) \ldots P(D_n | \theta) = \prod_i P(D_i | \theta)$$

- *Maximum Likelihood Estimation*

$$\theta^*_{ML} = \arg\max_\theta L(\theta | D)$$

When the observations are IID, the *Log-Likelihood* could ease computations:

$$\ell(\theta | D) = \log L(\theta | D) = \log \prod_i P(D_i | \theta) = \sum_i \log P(D_i | \theta)$$

$$\theta^*_{ML} = \arg\max_\theta \ell(\theta | D)$$

Example: coin tossing (*Bernoulli Trials*)

Test: tossing a coin X, not necessarily *fair*. ($X = 1$ head, $X = 0$ tail)

Model: $P(X = 1) = \theta$, $P(X = 0) = 1 - \theta$

Observations: a sequence $< D_1, D_2, \ldots , D_n >$
(i.e. $D = \{D_1=\{X_1 = 1\}, D_2=\{X_2 = 1\}, D_2=\{X_3 = 0\} \ldots\}$)

- *(Log-)Likelihood Function*

$$\ell(\theta|D) \;=\; \log P(D|\theta) = \log P(\{X_i\}|\theta) = \log \prod_i P(X_i|\theta) = \sum_i \log P(X_i|\theta)$$

Likelihood for P: (*Algebraic Follies!*)

$$P(X|\theta) = \theta^{[X=1]}(1-\theta)^{[X=0]} \quad \text{where:} \quad [X_i = v] = \begin{cases} 1 & se \quad X_i = v \\ 0 & se \quad X_i \neq v \end{cases}$$

$$\ell(\theta|D) \;=\; \sum_i \log\left(\theta^{[X_i=1]}(1-\theta)^{[X_i=0]}\right) = \log\theta \sum_i [X_i = 1] + \log(1-\theta)\sum_i [X_i = 0]$$

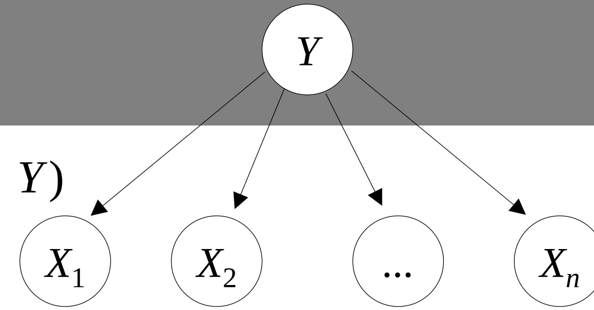
$$= \; N_{X=1}\log\theta + N_{X=0}\log(1-\theta)$$

(where $N_{X=1}$ is the number of $X_i = 1$ in the sequence D)

- *Maximum Likelihood Estimation*

$$\frac{\partial \ell}{\partial \theta} = \frac{N_{X=1}}{\theta} - \frac{N_{X=0}}{(1-\theta)} \qquad \frac{\partial \ell}{\partial \theta} = 0 \quad \Rightarrow \quad \theta^*_{ML} = \frac{N_{X=1}}{N_{X=1} + N_{X=0}}$$

Anti-spam filter



$$P(Y,\{X_i\}) = P(Y)\prod_{i=1}^{n} P(X_i \mid Y)$$

Model: the *conditional probability tables* in the graphical model

$$P(Y=k) = \pi_k\,,\quad P(X_i = j \mid Y=k) = \eta_{ijk}$$

Observations: a set of messages, with classification

$$D = \{D_1 = \{Y_1 = 1, X_{11} = 1, X_{12} = 1, \ldots, X_{1n} = 0\}, D_1 = \{Y_2 = 0, X_{21} = 0, X_{22} = 1, \ldots, X_{2n} = 1\}, \ldots\}$$

- *Likelihood Function*

Sequence of messages

$$L(\{\pi_k, \eta_{ijk}\} \mid D) = P(D \mid \theta) = P(\{D_m\} \mid \{\pi_k, \eta_{ijk}\}) = \prod_m P(D_m \mid \{\pi_k, \eta_{ijk}\})$$ (messages are IID)

$$= \prod_m P(\{Y_m = y_m, X_{mi} = x_{mi}\} \mid \{\pi_k, \eta_{ijk}\})$$

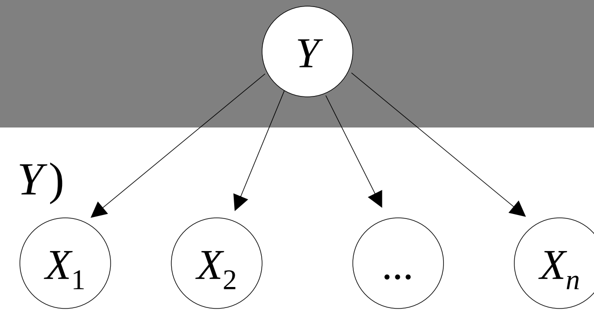
$$= \prod_m P(Y_m = y_m \mid \{\pi_k, \eta_{ijk}\})\, P(\{X_{mi} = x_{mi}\} \mid Y_m = y_m, \{\pi_k, \eta_{ijk}\})$$ (factorization)

$$= \prod_m P(Y_m = y_m \mid \{\pi_k\})\, P(\{X_{mi} = x_{mi}\} \mid Y_m = y_m, \{\eta_{ijk}\})$$ (cond. independence)

$$= \prod_m P(Y_m = y_m \mid \{\pi_k\}) \prod_i P(X_{mi} = x_{mi} \mid Y_m = y_m, \{\eta_{ijk}\})$$ ($<X_i \perp X_j, Y>$)

Anti-spam filter

$$P(Y,\{X_i\}) = P(Y)\prod_{i=1}^{n} P(X_i \mid Y)$$



- *Log-Likelihood Function*

$$\ell(\{\pi_k,\eta_{ijk}\} \mid D) = \sum_m \log P(Y_m = y_m \mid \{\pi_k\}) + \sum_m \sum_i \log P(X_{mi} = x_{mi} \mid Y_m = y_m, \{\eta_{ijk}\})$$

Alternative form for P*:*

$$P(Y = k \mid \{\pi_k\}) = \pi_k = \prod_k \pi_k^{[Y=k]}$$

(*Algebraic Follies!*)

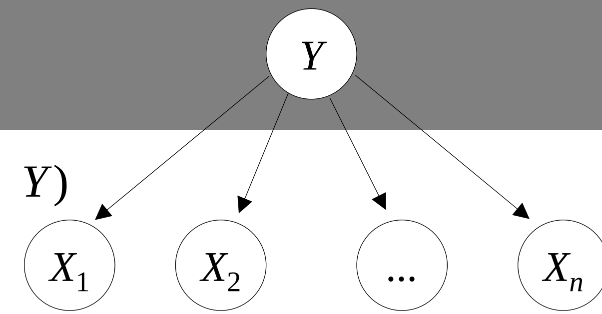
$$P(X_i = j \mid Y_m = k, \{\eta_{ijk}\}) = \eta_{ijk} = \prod_j \prod_k \eta_{ijk}^{[X_i=j][Y=k]}$$

$$\ell(\{\pi_k,\eta_{ijk}\} \mid D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \sum_m \sum_i \sum_j \sum_k [X_{mi} = j][Y_m = k] \log \eta_{ijk}$$

- *Maximum Likelihood Estimation*

Being both positive and depending on different variables, the two terms above can be optimized separately

Anti-spam filter



$$P(Y, \{X_i\}) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$

- *Maximum Likelihood Estimation*

$$\ell(\{\pi_k, \eta_{ijk}\} \mid D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \sum_m \sum_i \sum_j \sum_k [X_{mi} = j][Y_m = k] \log \eta_{ijk}$$

First term:

Lagrange multiplier

$$\ell^*(\{\pi_k\} \mid D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \lambda (1 - \sum_k \pi_k)$$

$$\frac{\partial \ell^*}{\partial \pi_k} = \frac{\sum_m [Y_m = k]}{\pi_k} - \lambda$$

number of messages in *D* classified as *k*

$$\frac{\partial \ell^*}{\partial \pi_k} = 0 \;\Rightarrow\; \pi_k = \frac{N_{Y=k}}{\lambda}$$

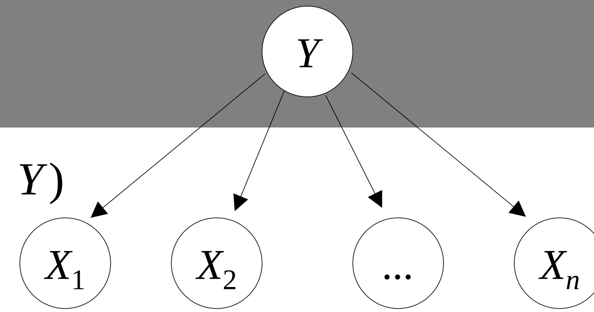
number of messages in *D*

$$\sum_k \pi_k = 1 \;\Rightarrow\; \sum_k \frac{N_{Y=k}}{\lambda} = 1 \;\Rightarrow\; \lambda = \sum_k N_{Y=k} = N_D$$

$$\pi_k^* = \frac{N_{Y=k}}{N_D}$$ (*Maximum Likelihood Estimator* of π_k)

Anti-spam filter

$$P(Y,\{X_i\}) = P(Y)\prod_{i=1}^{n} P(X_i \mid Y)$$



- *Maximum Likelihood Estimation*

$$\ell(\{\pi_k,\eta_{ijk}\}\mid D) = \sum_m\sum_k [Y_m = k]\log \pi_k + \sum_m\sum_i\sum_j\sum_k [X_{mi} = j][Y_m = k]\log \eta_{ijk}$$

Second term:

$$\ell^*(\{\eta_{ijk}\}\mid D) = \sum_m\sum_i\sum_j\sum_k [X_{mi} = j][Y_m = k]\log \eta_{ijk} + \sum_i\sum_k \lambda_{ik}(1-\sum_j \eta_{ijk})$$

$$\frac{\partial \ell^*}{\partial \eta_{ijk}} = \frac{\sum_m [X_{mi} = j][Y_m = k]}{\eta_{ijk}} - \lambda_{ik}$$

$$\frac{\partial \ell^*}{\partial \eta_{ijk}} = 0 \quad\Rightarrow\quad \eta_{ijk} = \frac{N_{X_i=j,Y=k}}{\lambda_{ik}}$$

$$\sum_j \eta_{ijk} = 1 \quad\Rightarrow\quad \sum_j \frac{N_{X_i=j,Y=k}}{\lambda_{ik}} = 1 \quad\Rightarrow\quad \lambda = \sum_j N_{X_i=j,Y=k} = N_{Y=k}$$

$$\eta^*_{ijk} = \frac{N_{X_i=j,Y=k}}{N_{Y=k}}$$

(*Maximum Likelihood Estimator* of η_{ijk})

Learning CPTs for a graphical model

As *Maximum Likelihood Estimation*

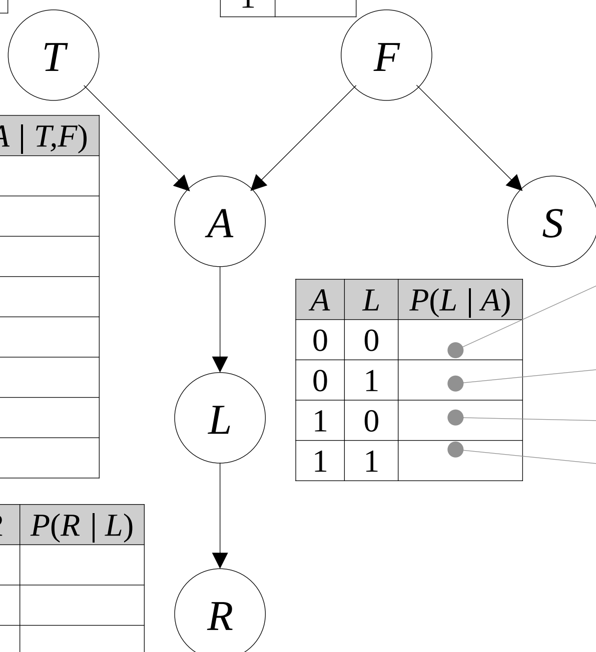
Model: the graphical model of the *fire alarm* example, with CPTs as parameters

Observations: sequence of sets di values, from <u>*completely observed*</u> situations

T	$P(T)$
0	$\frac{N_{T=0}}{N_D}$
1	$\frac{N_{T=1}}{N_D}$

F	$P(F)$
0	$\frac{N_{F=0}}{N_D}$
1	$\frac{N_{F=1}}{N_D}$

F	S	$P(S \mid F)$
0	0	$\frac{N_{S=0,F=0}}{N_{F=0}}$
0	1	$\frac{N_{S=1,F=0}}{N_{F=0}}$
1	0	$\frac{N_{S=0,F=1}}{N_{F=1}}$
1	1	$\frac{N_{S=1,F=1}}{N_{F=1}}$



T	F	A	$P(A \mid T,F)$
0	0	0	
0	0	1	
0	1	0	
0	1	1	
1	0	0	
1	0	1	
1	1	0	
1	1	1	

A	L	$P(L \mid A)$
0	0	$\frac{N_{L=0,A=0}}{N_{A=0}}$
0	1	$\frac{N_{L=1,A=0}}{N_{A=0}}$
1	0	$\frac{N_{L=0,A=1}}{N_{A=1}}$
1	1	$\frac{N_{L=1,A=1}}{N_{A=1}}$

L	R	$P(R \mid L)$
0	0	
0	1	
1	0	
1	1	

Bayesian learning

- *Maximum a Posteriori Estimation* (*MAP*)

  Instead of a *likelihood function,* the a posteriori probability is maximized

$$P(\theta | D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{P(D|\theta)P(\theta)}{\sum_{\theta} P(D|\theta)P(\theta)}$$

  Which is equivalent to optimize, w.r.t. θ:

$$P(D|\theta)\, P(\theta)$$

Advantages:

- Regularization: not all possible combinations of values might be present in D
- A formula for incremental learning:
  a priori terms could represent what was known *before* observations D

Problem:

- Which *prior* distribution $P(\theta)$?

Beta distribution

Gamma function (n integer > 0)

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

Beta function *(α and β integers > 0)*

$$B(\alpha,\beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$$

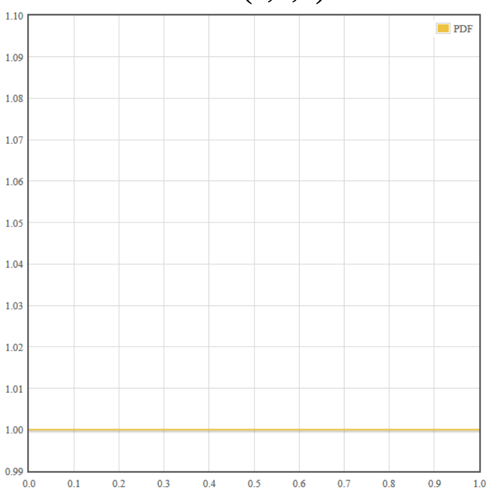
The definition is more complex when α and β are not integers (see Wikipedia)

- Beta probability density function (pdf) *(α and β integers > 0)*

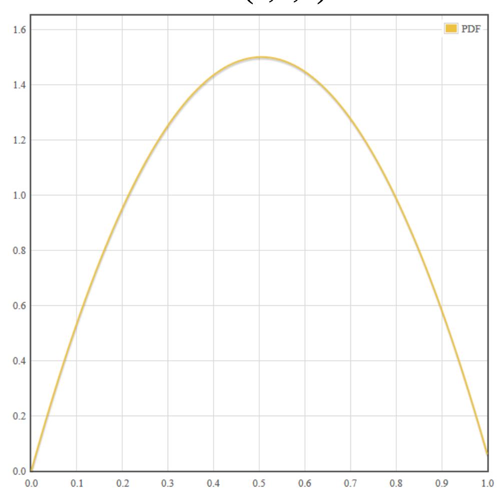
$$\text{Beta}(x;\alpha,\beta) := \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

The maximum occurs at: $x = \frac{\alpha-1}{\alpha-\beta-2}$

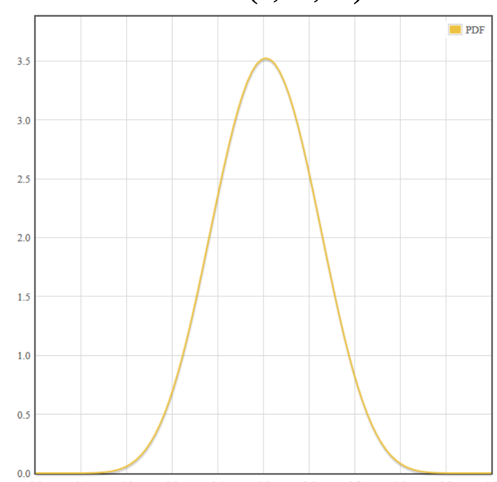
Beta(x;1,1)



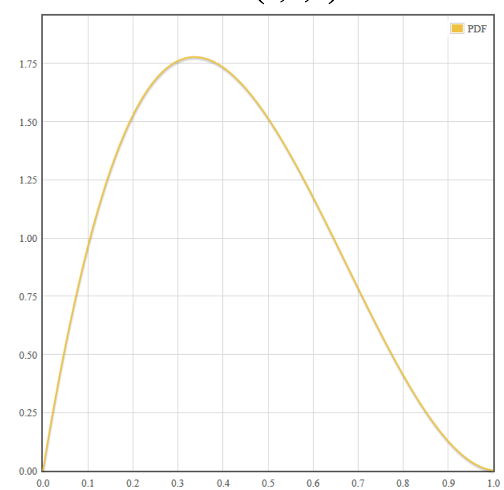
Beta(x;2,2)



Beta(x;10,10)



Beta(x;2,3)



Conjugate prior distributions

Coin tossing

$$P(D_i\,|\,\theta) \;=\; \theta^{[X_i=1]}(1-\theta)^{[X_i=0]}$$

(a.k.a. the Bernoulli distribution)

Likelihood (repeated experiments)

α_D and β_D are the result counts (i.e. heads and tails)

$$P(D\,|\,\theta) \;= P(\{D_i\}\,|\,\theta) \;=\; \prod_i P(D_i\,|\,\theta) \;= \theta^{\alpha_D}(1-\theta)^{\beta_D}$$

A posteriori probability with Beta prior

α_P and β_P are are the hyperparameters *of the prior*

$$P(D\,|\,\theta)\,P(\theta) = \theta^{\alpha_D}(1-\theta)^{\beta_D}\cdot \mathrm{Beta}(\theta;\alpha_P,\beta_P) = \theta^{\alpha_D}(1-\theta)^{\beta_D}\cdot\frac{\theta^{\alpha_P-1}(1-\theta)^{\beta_P-1}}{\mathrm{B}(\alpha_P,\beta_P)}$$

$$= \frac{\theta^{\alpha_D+\alpha_P-1}(1-\theta)^{\beta_D+\beta_P-1}}{\mathrm{B}(\alpha_P,\beta_P)} = \frac{\mathrm{B}(\alpha_D+\alpha_P,\beta_D+\beta_P)}{\mathrm{B}(\alpha_P,\beta_P)}\cdot \mathrm{Beta}(\theta;\alpha_D+\alpha_P,\beta_D+\beta_P)$$

this factor is a positive constant (for θ)

Moral:

$$P(D\,|\,\theta)\,P(\theta) \;\propto\; \mathrm{Beta}(\theta;\alpha_D+\alpha_P,\beta_D+\beta_P)$$

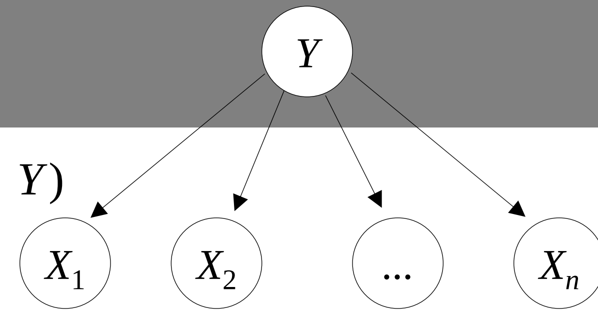
Therefore

$$\theta^*_{MAP} = \arg\max_\theta \mathrm{Beta}(\theta;\alpha_D+\alpha_P,\beta_D+\beta_P) = \frac{\alpha_D+\alpha_P-1}{\alpha_D+\alpha_P+\beta_D+\beta_P-2}$$

It is the same result as MLE but with the addition of $\alpha_P+\beta_P-2$ *pseudo-observations*

Being a ***conjugate prior*** $P(\theta)$ of a distribution $P(D\,|\,\theta)$ *in the above sense*
means that the posterior $P(D\,|\,\theta)\,P(\theta)$ is in *the same family* of $P(\theta)$

Anti-spam filter



$$P(Y,\{X_i\}) = P(Y)\prod_{i=1}^{n} P(X_i \mid Y)$$

- *Maximum a Posteriori (MAP) Estimation*

  The adapted computations for:

  $$\theta^*_{MAP} = \arg\max_\theta P(D\,|\,\theta)\, P(\theta)$$

  yield:

  $$\pi^*_k = \frac{\alpha_k + N_{Y=k} - 1}{\alpha_k + \beta_k + N_D - 2}$$ (*MAP Estimator* of π_k)

  $$\eta^*_{ijk} = \frac{\alpha_{ijk} + N_{X_i=j,Y=k} - 1}{\alpha_{ijk} + \beta_{ijk} + N_{Y=k} - 2}$$ (*MAP Estimator* of η_{ijk})

  where the

  $$\alpha_k, \beta_k, \alpha_{ijk}, \beta_{ijk}$$

  are the *hyperparameters* of the prior distribution representing the *pseudo-observations* made *before* the arrival of new, actual observations D