Artificial Intelligence

An Advanced Course About Foundations

UNIVERSITÀ DI PAVIA

Probabilistic Reasoning: Supervised Learning

Marco Piastra

Machine Learning

Types of machine learning problems

Consider a number of *observations* (i.e. a *dataset*) made by an agent

$\{d^{(1)}, d^{(2)}, ..., d^{(N)}\}$

Also called 'data items' (see next slide)

- ***Supervised learning***

  Learning form *complete* observations: each of the observations $\{d^{(1)}, d^{(2)}, ..., d^{(N)}\}$ include values for all random variables in the model

  The objective is learning a *distribution* P

- ***Unsupervised learning***

  Learning from *incomplete* observations: observations $\{d^{(1)}, d^{(2)}, ..., d^{(N)}\}$ do not necessarily include values for all the random variables in the model

  The objective is learning a *distribution* P

- ***Reinforcement learning***

  The observations $\{d^{(1)}, d^{(2)}, ..., d^{(N)}\}$ are *states* o *situations*,
  at each ***state*** $S^{(i)}$ the agent performs an ***action*** a_i that produces a new ***state*** $S^{(i+1)}$ a ***reward*** r_{i+1}.

  The objective is learning a *distribution* π over *actions* in each *state*
  which describes the ***policy*** that the agent will follow

ML = Representation + Evaluation + Optimization

- **Representation**

  The objective is learning a specific distribution

  $$P(\{X_r\};\theta)$$

  where $\{X_r\}$ are all the random variables of interest and θ is a *set* of parameters

  Which kind of distribution (i.e. the *model* or also the *learner*) do we select?

  Example: assume we select the anti-spam filter (i.e. Naïve Bayesian Classifier) as the model the parameters in such case are the numerical probabilities in the CPTs

- **Evaluation**

  Given a dataset D, how well does a specific set of parameter values $\hat{\theta}$ make the distribution P fit the dataset?

  An estimator, i.e. a scoring function of some sort, must be selected

- **Optimization**

  How can we find the optimal set of parameter values θ^* with respect to the *estimator* of choice?

  In general, this is an optimization problem

Observations and Datasets

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

The outcome of a particular experiment can be represented by a vector of *random variables*

Example: if the model includes two random variables $\{X, Y\}$
the N outcomes of the experiments are $d^{(1)} = (X^{(1)}, Y^{(1)}), \ \ldots \ , d^{(N)} = (X^{(N)}, Y^{(N)})$

Why representing data as random variables?

- Actual values are not deterministic but drawn from some distribution
- The distribution reflects our uncertainty or variability in the process
- Actual data might be known or *partially* known

More in general, each observation is described by a *data item*

$$d^{(i)} = (X_1^{(i)}, \ldots, X_n^{(i)})$$

In general, each observation is a vector of random variables

A *dataset* is a collection of data items

$$D := \{d^{(i)}\}_{i=1}^{N}$$

N could be very large

Independent and Identically Distributed (IID)

- **Independent observations, same probability distribution**

Random variables in a set $\{X^{(1)}, X^{(2)}, \ldots, X^{(N)}\}$
are *Independent and Identically Distributed* (IID) iff:

$$\langle X^{(i)} \perp X^{(j)} \rangle, \ \forall i \neq j \qquad \text{(independence)}$$

$$P(X^{(i)}) = P(X^{(j)}), \ \forall i \neq j \qquad \text{(identical distribution)}$$

CAUTION: *Being* IID *is not an obvious property of observations*
e.g. same measurements on different patients may be IID,
but different measurements over time on the same patient are not IID

Observations in a *dataset* $\{d^{(1)}, d^{(2)}, \ldots, d^{(N)}\}$
are *Independent and Identically Distributed* (IID) iff:

$$\langle d^{(i)} \perp d^{(j)} \rangle, \ \forall i \neq j \implies \langle (X_1^{(i)}, \ldots, X_n^{(i)}) \perp (X_1^{(j)}, \ldots, X_n^{(j)}) \rangle$$

$$P(d^{(i)}) = P(d^{(j)}), \ \forall i \neq j, \implies P(X_1^{(i)}, \ldots, X_n^{(i)}) = P(X_1^{(j)}, \ldots, X_n^{(j)})$$

In an IID dataset,
observations,
not individual variables,
must be independent
and identically distributed

Maximum Likelihood Estimator (MLE)

Evaluation: Likelihood

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics

- ***Statistical (inductive) inference*: probability measure from data**

  - A probabilistic model $P(D; \theta)$ where θ are the *parameters*

    Parameters θ: a set of values that characterizes P *completely* (once θ is defined, P is also defined).
  - A dataset of IID observations (*data items*) $D = \{d^{(1)}, d^{(2)}, ..., d^{(N)}\}$

Likelihood function

$$L(\theta, D) := P(D; \theta) = P(d^{(1)}, \ldots, d^{(N)};\ \theta)$$

Likelihood is a scalar function

The probability of the entire dataset when the parameters θ have a specific value

When the observations $\{d^{(1)}, \ldots, d^{(N)}\}$ are IID:

$$P(d^{(1)}, \ldots, d^{(N)};\ \theta) = \prod_{i=1}^{N} P(d^{(i)};\ \theta)$$

Maximum Likelihood Estimator (MLE)

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics

- ***Statistical (inductive) inference*: probability measure from data**

  - A probabilistic model $P(D; \theta)$ where θ are the *parameters*

    Parameters θ: a <u>set</u> of values that characterizes P *completely* (once θ is defined, P is also defined).
  - A dataset of IID observations (*data items*) $D = \{d^{(1)}, d^{(2)}, ..., d^{(N)}\}$

Maximum Likelihood Estimator (MLE)

$$\theta^*_{ML} := \underset{\theta}{\operatorname{argmax}}\ L(\theta, D)$$

The set of parameters θ^* that maximize the probability of <u>the entire dataset</u>

When the observations $\{d^{(1)}, \dots, d^{(N)}\}$ are IID:

$$P(d^{(1)}, \dots, d^{(N)};\ \theta) = \prod_{i=1}^{N} P(d^{(i)};\ \theta)$$

Maximum Likelihood Estimator (MLE)

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics

- ***Statistical (inductive) inference*: probability measure from data**

  - A probabilistic model $P(D; \theta)$ where θ are the *parameters*

    Parameters θ: a <u>set</u> of values that characterizes P *completely* (once θ is defined, P is also defined).
  - A dataset of IID observations (*data items*) $D = \{d^{(1)}, d^{(2)}, ..., d^{(N)}\}$

Maximum Likelihood Estimator (MLE, log-likelihood)

$$\ell(\theta, D) := \log L(\theta, D)$$

$$\theta^*_{ML} = \underset{\theta}{\operatorname{argmax}}\ \ell(\theta, D)$$

true because log *is monotonically increasing*

When the observations $\{d^{(1)}, \dots, d^{(N)}\}$ are IID:

$$\ell(\theta, D) = \log L(\theta, D) = \log \prod_{i=1}^{N} P(d^{(i)}; \theta) = \sum_{i=1}^{N} \log P(d^{(i)}; \theta)$$

Maximum Likelihood Estimator (MLE)

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics

- ***A Success Story***

  Desirable statistical properties of the *Maximum Likelihood Estimator*:

  - *Consistency*: Estimates converge to the true parameters as sample size grows.
  - *Efficiency*: Achieves minimum variance among unbiased estimators (asymptotically)

    unbiased estimator: one such that $\mathbb{E}_{\mathcal{D}}[\hat{\theta}] = \theta^*$
    (the expectation over all possible datasets)

  - *Sufficiency*: Uses all information in the data relevant to the parameters

Finding θ^*_{ML} *is an optimization problem*

MLE as Optimization (The Analytical Way)

Example: coin tossing (*Bernoulli Trials*)

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

Parameter (*just one*): $P(X=1)=\theta, \quad P(X=0)=1-\theta$

Observations: a sequence of experimental outcomes

$$D=\{d^{(1)}, \ldots, d^{(N)}\}=\{(X^{(1)}=x^{(1)}), \ldots, (X^{(N)}=x^{(N)})\}$$

- Binomial distribution

  The probability of obtaining $N_{X=1}$ times 'head' in a sequence of N trials

$$\binom{N}{k} := \frac{N!}{k!\,(N-k)!} \quad \textit{binomial coefficient}$$

$$P(D;\theta)=\binom{N}{N_{X=1}}\prod_i P(X^{(i)};\theta)=\binom{N}{N_{X=1}}P(X=1;\theta)^{N_{X=1}}\;P(X=0;\theta)^{N_{X=0}}$$

$N_{X=1}$ is the number of X=1 (i.e. heads) in a sequence of N trials

$$=\binom{N}{N_{X=1}}\theta^{N_{X=1}}(1-\theta)^{N_{X=0}}$$

In this case, it is assumed to be the likelihood of $\{d^{(1)}, \ldots, d^{(N)}\}$ *given the parameter* θ

Example: coin tossing (*Bernoulli Trials*)

- *(Log-)Likelihood Function*

IID assumption

$$\ell(\theta, D) = \log P(D;\theta) = \log P(\{X^{(i)}\};\ \theta) = \log \binom{N}{N_{X=1}} \prod_i P(X^{(i)};\ \theta) = \log \binom{N}{N_{X=1}} + \sum_i \log P(X^{(i)};\ \theta)$$

Rewrite $P(X\,;\,\theta)$ as:

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

Also called *indicator function*

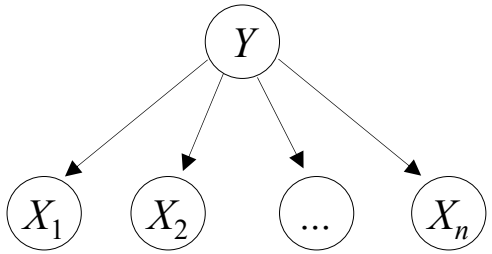
$$\begin{aligned} \ell(\theta, D) &= \log \binom{N}{N_{X=1}} + \sum_i \log \left(\theta^{\,[X^{(i)}=1]}\ (1\ -\ \theta)^{[X^{(i)}=0]} \right) \\ &= \log \binom{N}{N_{X=1}}\ +\ \log\theta \sum_i [X^{(i)}=1]\ +\ \log(1-\theta) \sum_i [X^{(i)}=0] \\ &= \log \binom{N}{N_{X=1}}\ +\ N_{X=1}\log\theta\ +\ N_{X=0}\log(1-\theta) \end{aligned}$$

- *Maximum Likelihood Estimation*

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

Naïve Bayesian Classifier

$$P(Y, X_1, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



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

$$\theta := \{\theta_y,\ \{\theta_{x_i,y}\}_{i=1}^{n}\}_{x_i,y\in\{0,1\}}, \quad \theta_y := P(Y = y), \quad \theta_{x_i,y} := P(X_i = x_i \mid Y = y)$$

Observations: a set of vectors representing email messages *with classification*

$$D = \left\{d^{(m)}\right\}_{m=1}^{N} = \left\{Y^{(m)} = y^{(m)}, \{X_i^{(m)} = x_i^{(m)}\}_{i=1}^{n}\right\}_{m=1}^{N}$$

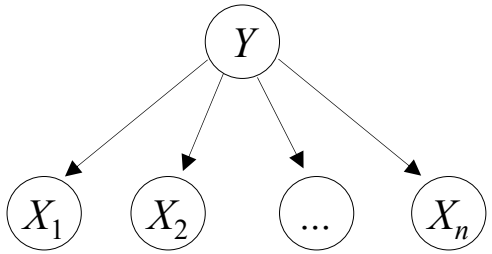
- *Likelihood Function*

$$L(\theta, D) \ = \ P(D;\ \theta) \ = \ P(D;\ \{\theta_y, \{\theta_{x_i,y}\}\}) \ = \ \prod_{m=1}^{N} P(d^{(m)};\ \{\theta_y, \{\theta_{x_i,y}\}\}) \quad \text{(data items are IID)}$$

$$\begin{aligned}
&= \prod_m P(\{Y^{(m)} = y^{(m)}, \{X_i^{(m)} = {x_i}^{(m)}\}\};\ \{\theta_y, \{\theta_{x_i,y}\}\}) && \text{(factorization)}\\
&= \prod_m P(Y^{(m)} = y^{(m)};\ \{\theta_y, \{\theta_{x_i,y}\}\})\ P(\{{X_i}^{(m)} = {x_i}^{(m)}\} \mid Y^{(m)} = y^{(m)};\ \{\theta_y, \{\theta_{x_i,y}\}\}) && \text{(cond. independence)}\\
&= \prod_m P(Y^{(m)} = y^{(m)};\ \theta_y)\ P(\{{X_i}^{(m)} = {x_i}^{(m)}\} \mid Y^{(m)} = y^{(m)};\ \{\theta_{x_i,y}\}) && (<X_i \perp X_j \mid Y>)\\
&= \prod_m P(Y^{(m)} = y^{(m)};\ \theta_y)\ \prod_i P({X_i}^{(m)} = {x_i}^{(m)} \mid Y^{(m)} = y^{(m)};\ \{\theta_{x_i,y}\})
\end{aligned}$$

Naïve Bayesian Classifier

$$P(Y, X_1, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



- *Log-Likelihood Function*

$$\ell(\{\theta_y, \{\theta_{x_i,y}\}\}, D) = \sum_m \log P(Y^{(m)} = y^{(m)};\ \theta_y) + \sum_m \sum_i \log P(X_i{}^{(m)} = x_i{}^{(m)} \mid Y^{(m)} = y^{(m)};\ \{\theta_{x_i,y}\})$$

Alternative form for P*:* *(rewritten using indicator functions)*

$$P(Y = y \mid \theta_y) = \prod_y \theta_y{}^{[Y=y]}$$

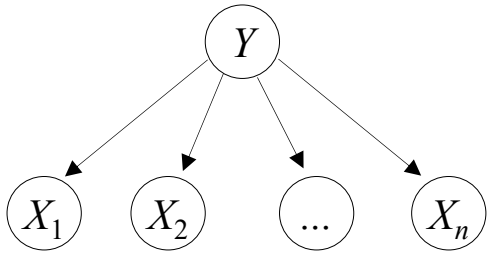
$$P(X_i = x_i \mid Y = y;\ \{\theta_{x_i,y}\}) = \prod_{x_i} \prod_y \theta_{x_i,y}{}^{[X_i=x_i][Y=y]}$$

$$\ell(\{\theta_y, \{\theta_{x_i,y}\}\}, D) = \sum_m \sum_y [Y^{(m)} = y] \log \theta_y + \sum_m \sum_{x_i} \sum_y [X_i{}^{(m)} = x_i][Y^{(m)} = y] \log \theta_{x_i,y}$$

Since they depend on different variables,
these two terms above can be optimized separately

Naïve Bayesian Classifier

$$P(Y, X_1, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



- *Maximum Likelihood Estimation*

$$\ell(\{\theta_y, \{\theta_{x_i,y}\}\}, D) = \sum_m \sum_y [Y^{(m)} = y] \log \theta_y + \sum_m \sum_{x_i} \sum_y [X_i{}^{(m)} = x_i][Y^{(m)} = y] \log \theta_{x_i,y}$$

Optimizing the first term:

Lagrange multiplier

$$\ell^*(\theta_y, D) = \sum_m \sum_y [Y^{(m)} = y] \log \theta_y + \lambda \left(1 - \sum_y \theta_y\right)$$

$$\frac{\partial \ell^*}{\partial \theta_y} = \frac{\sum_m [Y^{(m)} = y]}{\theta_y} - \lambda$$

number of messages in D classified as k

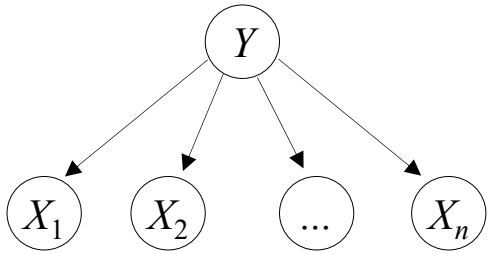
$$\frac{\partial \ell^*}{\partial \theta_y} = 0 \quad \Rightarrow \quad \theta_y = \frac{N_{Y=y}}{\lambda}$$

$$\sum_y \theta_y = 1 \quad \Rightarrow \quad \sum_y \frac{N_{Y=y}}{\lambda} = 1 \quad \Rightarrow \quad \lambda = \sum_y N_{Y=y} = N$$

$$\theta_y^* = \frac{N_{Y=y}}{N} \qquad \textit{(Maximum Likelihood Estimator)}$$

Naïve Bayesian Classifier

$$P(Y, X_1, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



- *Maximum Likelihood Estimation*

$$\ell(\{\theta_y, \{\theta_{x_i,y}\}\}, D) = \sum_m \sum_y [Y^{(m)} = y] \log \theta_y + \sum_m \sum_{x_i} \sum_y [X_i{}^{(m)} = x_i][Y^{(m)} = y] \log \theta_{x_i,y}$$

Optimizing the second term:

Lagrange multipliers

$$\ell^*(\{\theta_{x_i,y}\}, D) = \sum_m \sum_i \sum_{x_i} \sum_y [X_i{}^{(m)} = x_i][Y^{(m)} = y] \log \theta_{x_i,y} + \sum_i \sum_y \lambda_{x_i,y} \left(1 - \sum_{x_i} \theta_{x_i,y}\right)$$

$$\frac{\partial \ell^*}{\partial \theta_{x_i,y}} = \frac{\sum_m [X_i{}^{(m)} = x_i][Y^{(m)} = y]}{\theta_{x_i,y}} - \lambda_{x_i,y}$$

$$\frac{\partial \ell^*}{\partial \theta_{x_i,y}} = 0 \quad \Rightarrow \quad \theta_{x_i,y} = \frac{N_{X_i=x_i,\, Y=y}}{\lambda_{x_i,y}}$$

$$\sum_{x_i} \theta_{x_i,y} = 1 \quad \Rightarrow \quad \sum_{x_i} \frac{N_{X_i=x_i,\, Y=y}}{\lambda_{x_i,y}} = 1 \quad \Rightarrow \quad \lambda_{x_i,y} = \sum_{x_i} N_{X_i=x_i, Y=y} = N_{Y=y}$$

$$\theta^*_{x_i,y} = \frac{N_{X_i=x_i,Y=y}}{N_{Y=y}} \qquad \textit{(Maximum Likelihood Estimator)}$$

MLE and discrete probability distributions

- ***MLE as relative frequencies***

  The *maximum likelihood estimator* (MLE) for any discrete probability distribution over a dataset is computed by determining the proportion of times each outcome occurs in the dataset

  This method uses the *relative frequencies* of the outcomes to provide the best estimate of their probabilities

MLE and discrete probability distributions

- ***MLE as relative frequencies (in subspaces)***

  The *maximum likelihood estimator* (MLE) for any discrete probability distribution over a dataset is computed by determining the proportion of times each outcome occurs in the dataset

  This method uses the *relative frequencies* of the outcomes to provide the best estimate of their probabilities

  When considering *conditional probabilities*, the MLE can be applied within specific subspaces of the dataset by focusing on the subset of the data where that condition is met

MLE for Graphical Models: A Fundamental Rule*

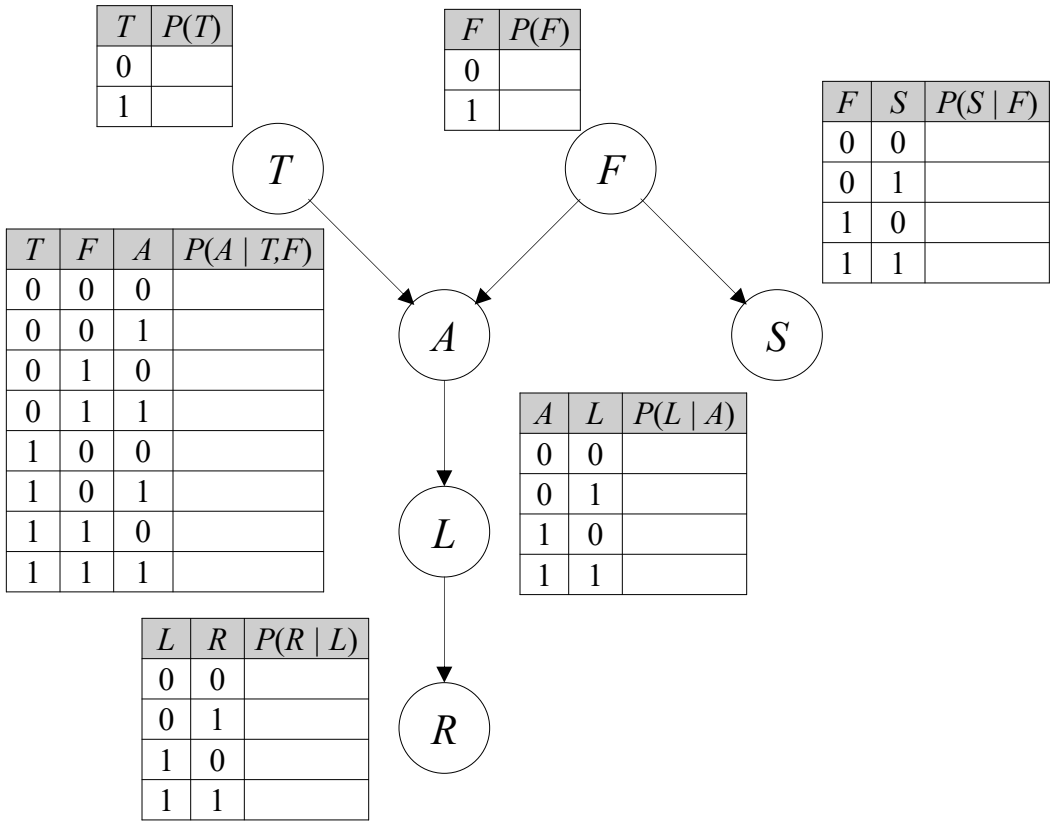
(*) In a finitary setting

Learning CPTs for a graphical model via MLE

Model: random variables plus the graph of dependencies

Observations: dataset of values, from *completely observed* outcomes

Parameters (to be determined): all conditional probabilities *(i.e. all CPTs)*

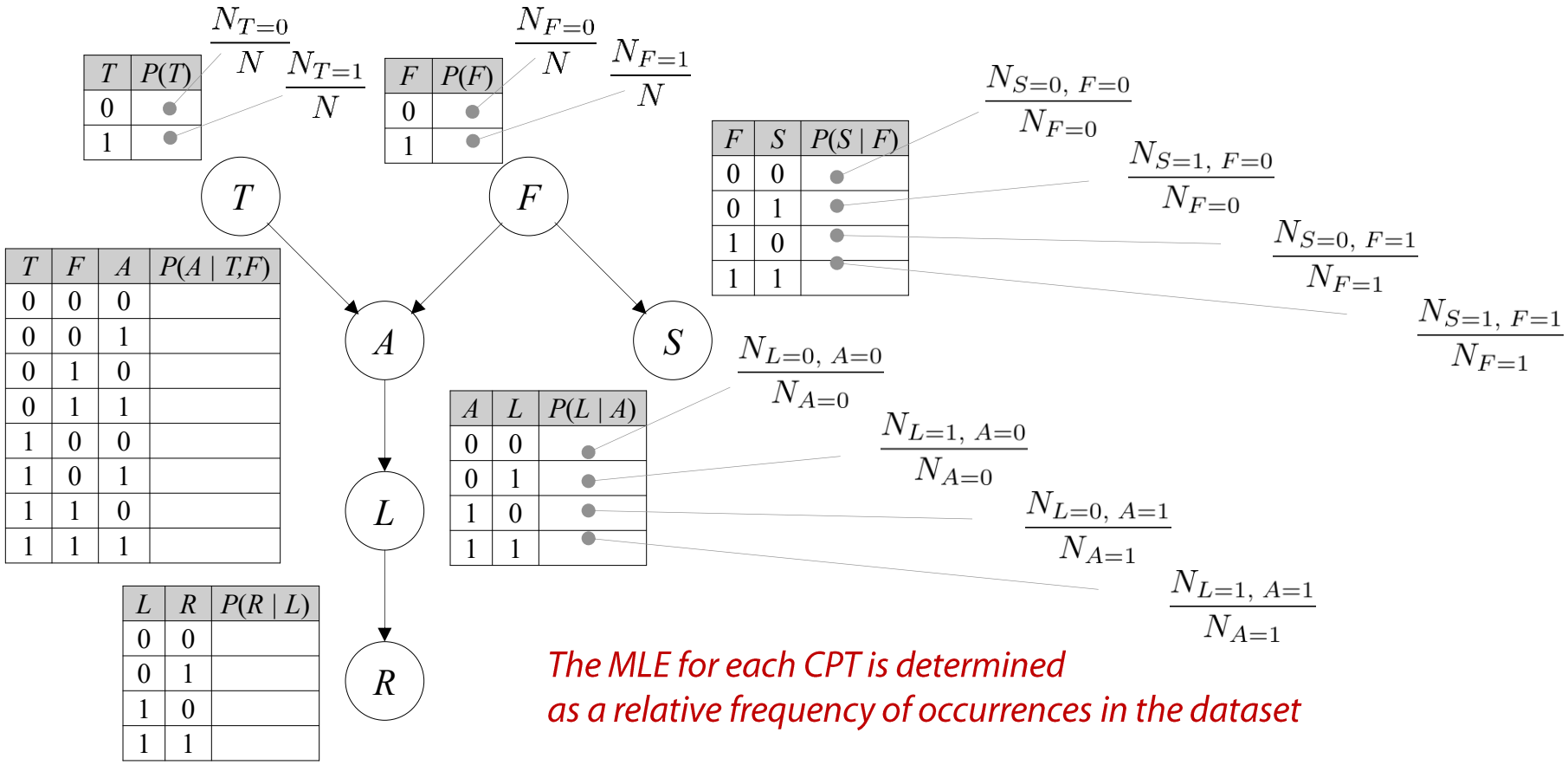


Learning CPTs for a graphical model via MLE

Model: random variables plus the graph of dependencies

Observations: dataset of values, from *completely observed* outcomes

Parameters (to be determined): all conditional probabilities *(i.e. all CPTs)*



T	P(T)
0	
1	

F	P(F)
0	
1	

F	S	P(S | F)
0	0	
0	1	
1	0	
1	1	

T	F	A	P(A | 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	

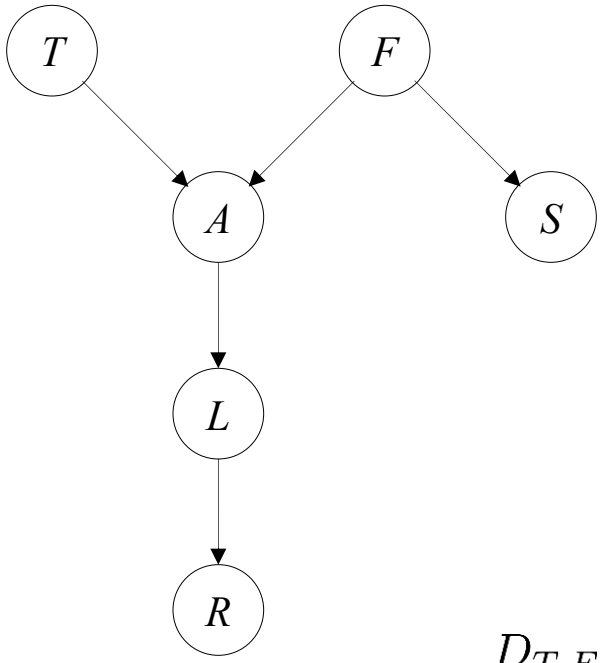
A	L	P(L | A)
0	0	
0	1	
1	0	
1	1	

L	R	P(R | L)
0	0	
0	1	
1	0	
1	1	

The MLE for each CPT is determined
as a relative frequency of occurrences in the dataset

Learning CPTs for a graphical model via MLE

More in general:
The MLE of a (directed) graphical model is the MLE of each node
(in each corresponding observation subset)



$$\theta^*_{ML} := \text{argmax}_\theta \; P(D \mid \theta)$$

$$\theta = \{\theta_T, \theta_F, \theta_{S|F}, \theta_{A|S,F}, \theta_{L|A}, \theta_{R|L}\}$$

$$\theta^*_T := \text{argmax}_{\theta_T} \; P(D;\; \theta_T)$$
$$\theta^*_F := \text{argmax}_{\theta_F} \; P(D;\; \theta_F)$$
$$\theta^*_{S|F} := \text{argmax}_{\theta_{S|F}} \; P(D_F;\; \theta_{S|F})$$
$$\theta^*_{A|T,F} := \text{argmax}_{\theta_{A|T,F}} \; P(D_{T,F};\; \theta_{A|T,F})$$
$$\theta^*_{L|A} := \text{argmax}_{\theta_{L|A}} \; P(D_A;\; \theta_{L|A})$$
$$\theta^*_{R|L} := \text{argmax}_{\theta_{R|L}} \; P(D_L;\; \theta_{R|L})$$

$D_{T,F}$ *denotes the subset of complete observation in which the random variables T, F have the corresponding values*

Bayesian Learning: Maximum a Posteriori (MAP) estimator

Bayesian learning

Instead of a *likelihood function*, the a *posteriori probability* is maximized w.r.t. θ

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

This is equivalent to maximize:

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

Maximum a Posteriori Estimation (*MAP*)

$$\theta^*_{MAP} := \underset{\theta}{\operatorname{argmax}}\; P(D \,|\theta)\; P(\theta)$$

Motivation:

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

Problem:

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

Beta distribution

See demo at https://www.acsu.buffalo.edu/~adamcunn/probability/beta.html

Gamma function (n integer > 0)

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

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

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

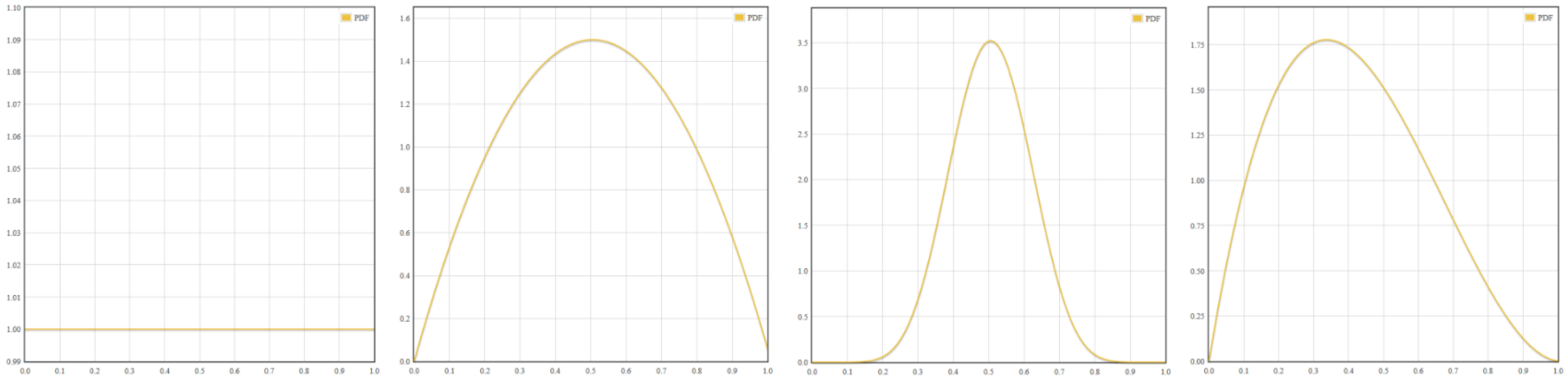
The definition is more complex when α and β are not integers (see Wikipedia)

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

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

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

Beta(θ; 1,1) Beta(θ; 2,2) Beta(θ; 10,10) Beta(θ; 2,3)



Conjugate prior distributions

Coin tossing (Binomial Distribution)

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

$$P(D\,|\theta) \;=\; \binom{\alpha_D+\beta_D}{\alpha_D}\prod_i P(X_i \mid \theta) \;=\; \binom{\alpha_D+\beta_D}{\alpha_D}\,\theta^{\alpha_D}\,(1-\theta)^{\beta_D}$$

A posteriori probability with Beta prior

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

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

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

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

$$= \binom{\alpha_D+\beta_D}{\alpha_D}\frac{\mathrm{B}(\alpha_D+\alpha_P,\beta_D+\beta_P)}{\mathrm{B}(\alpha_P,\beta_P)}\mathrm{Beta}(\theta\,;\alpha_D+\alpha_P,\beta_D+\beta_P)$$

this is a positive constant (no dependency on θ*)*

Conjugate prior distributions

Coin tossing (Binomial Distribution)

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

$$P(D\,|\theta) \;=\; \binom{\alpha_D+\beta_D}{\alpha_D}\prod_i P(X_i \mid \theta) \;=\; \binom{\alpha_D+\beta_D}{\alpha_D}\,\theta^{\alpha_D}\,(1-\theta)^{\beta_D}$$

A posteriori probability with Beta prior

$$P(D\,|\theta)P(\theta) = \binom{\alpha_D+\beta_D}{\alpha_D}\frac{\mathrm{B}(\alpha_D+\alpha_P,\beta_D+\beta_P)}{\mathrm{B}(\alpha_P,\beta_P)}\mathrm{Beta}(\theta\,;\alpha_D+\alpha_P,\beta_D+\beta_P)$$

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

"is proportional to"

Maximization:

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

This is the same as MLE but with the addition of $(\alpha_P-1)+(\beta_P-1)$ *pseudo-observations*

Conjugate prior distributions

Coin tossing (i.e. a specific observation i)

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

Likelihood (of a dataset)

$$P(D \,|\theta) \;=\; \binom{N}{N_{X=1}} \prod_i P(X_i \,|\theta) \;=\; \binom{N}{N_{X=1}} \theta^{N_{X=1}}(1-\theta)^{N_{X=0}}$$

A posteriori probability with Beta prior

$$P(D \,|\theta)P(\theta) \;\propto\; \mathrm{Beta}(\theta \,; N_{X=1} + \alpha_P,\; N_{X=0} + \beta_P)$$

"is proportional to"

Therefore:

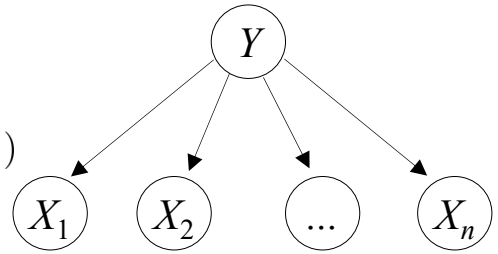
$$\theta^*_{MAP} \;=\; \mathrm{argmax}_\theta \,\mathrm{Beta}(\theta \,; N_{X=1} + \alpha_P,\; N_{X=0} + \beta_P) \;=\; \frac{N_{X=1} + \alpha_P - 1}{N + \alpha_P + \beta_P - 2}$$

which is the same as MLE but with the addition of $\alpha_P + \beta_P$ *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 the prior $P(\theta)$

Anti-spam filter

$$P(Y, X_1, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid X_{i-1})$$



- *Maximum a Posteriori (MAP) Estimation*

  The adapted computations for:

$$\theta^*_{MAP} := \text{argmax}_\theta \; P(D \mid \theta) \, P(\theta)$$

  yield:

$$\theta^*_k = \frac{\alpha_k + N_{Y=k} - 1}{\alpha_k + \beta_k + N - 2}$$ (*MAP Estimator* of Θ_k)

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

  where:

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

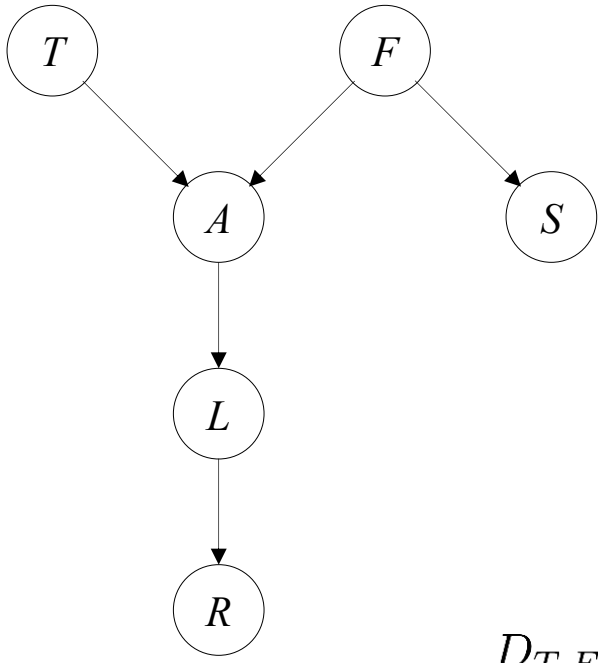
  are the *hyperparameters* of the prior distribution representing the *pseudo-observations* made *before* the arrival of new, actual observations D

Bayesian Learning: MAP for Graphical Models

Learning CPTs for a graphical model

As *Maximum a Posteriori Estimation*

More in general:
The MAP of a (directed) graphical model is the MAP of each node
(in each corresponding observation subset)



$$\theta^*_{MAP} := \text{argmax}_\theta \; P(D \mid \theta) \; P(\theta)$$

$$\theta = \{\theta_T, \theta_F, \theta_{S|F}, \theta_{A|S,F}, \theta_{L|A}, \theta_{R|L}\}$$

$$\theta^*_T := \text{argmax}_{\theta_T} \; P(D \mid \theta_T) \; P(\theta_T)$$
$$\theta^*_F := \text{argmax}_{\theta_F} \; P(D \mid \theta_F) \; P(\theta_F)$$
$$\theta^*_{S|F} := \text{argmax}_{\theta_{S|F}} \; P(D_F \mid \theta_{S|F}) \; P(\theta_{S|F})$$
$$\theta^*_{A|T,F} := \text{argmax}_{\theta_{A|T,F}} \; P(D_{T,F} \mid \theta_{A|T,F}) \; P(\theta_{A|T,F})$$
$$\theta^*_{L|A} := \text{argmax}_{\theta_{L|A}} \; P(D_A \mid \theta_{L|A}) \; P(\theta_{L|A})$$
$$\theta^*_{R|L} := \text{argmax}_{\theta_{R|L}} \; P(D_L \mid \theta_{R|L}) \; P(\theta_{R|L})$$

$D_{T,F}$ *denotes the subset of complete observation in which the random variables T, F have the corresponding value*