10-701 Fall 2017 Recitation 2

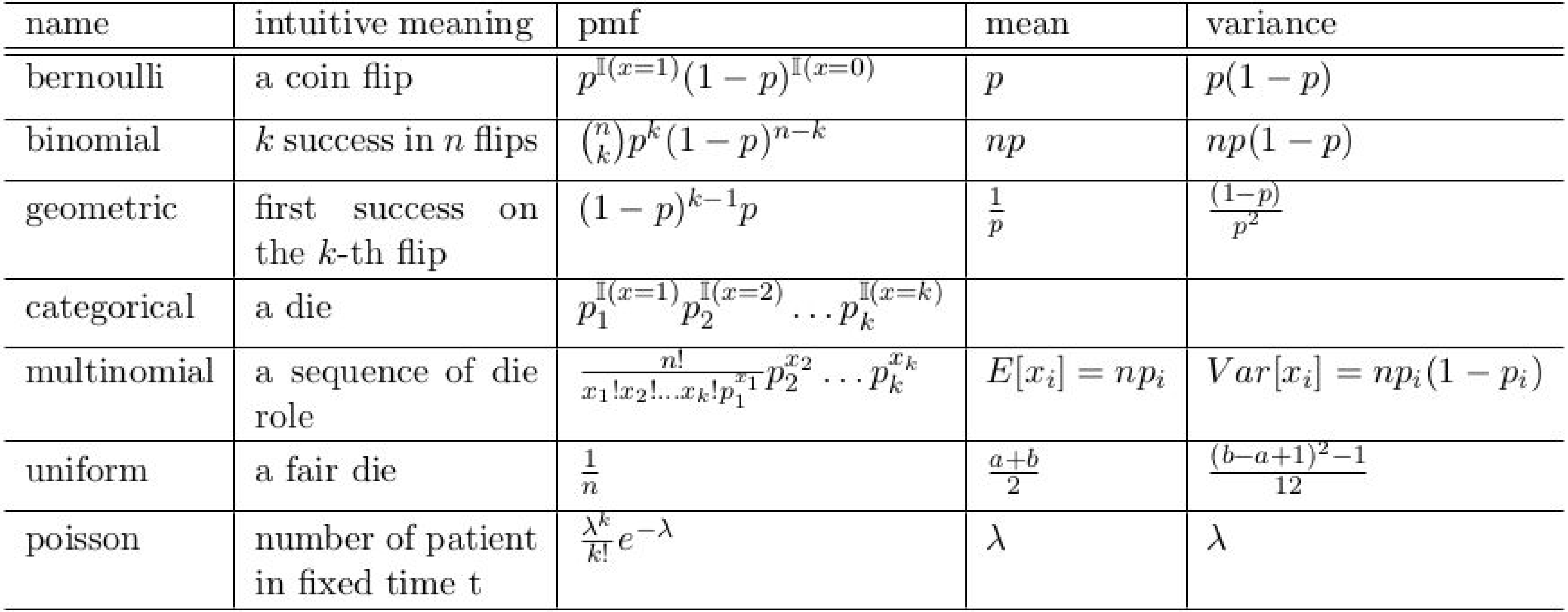
Yujie, Jessica, Akash

Probability Review

Theory on basic probability and expectation

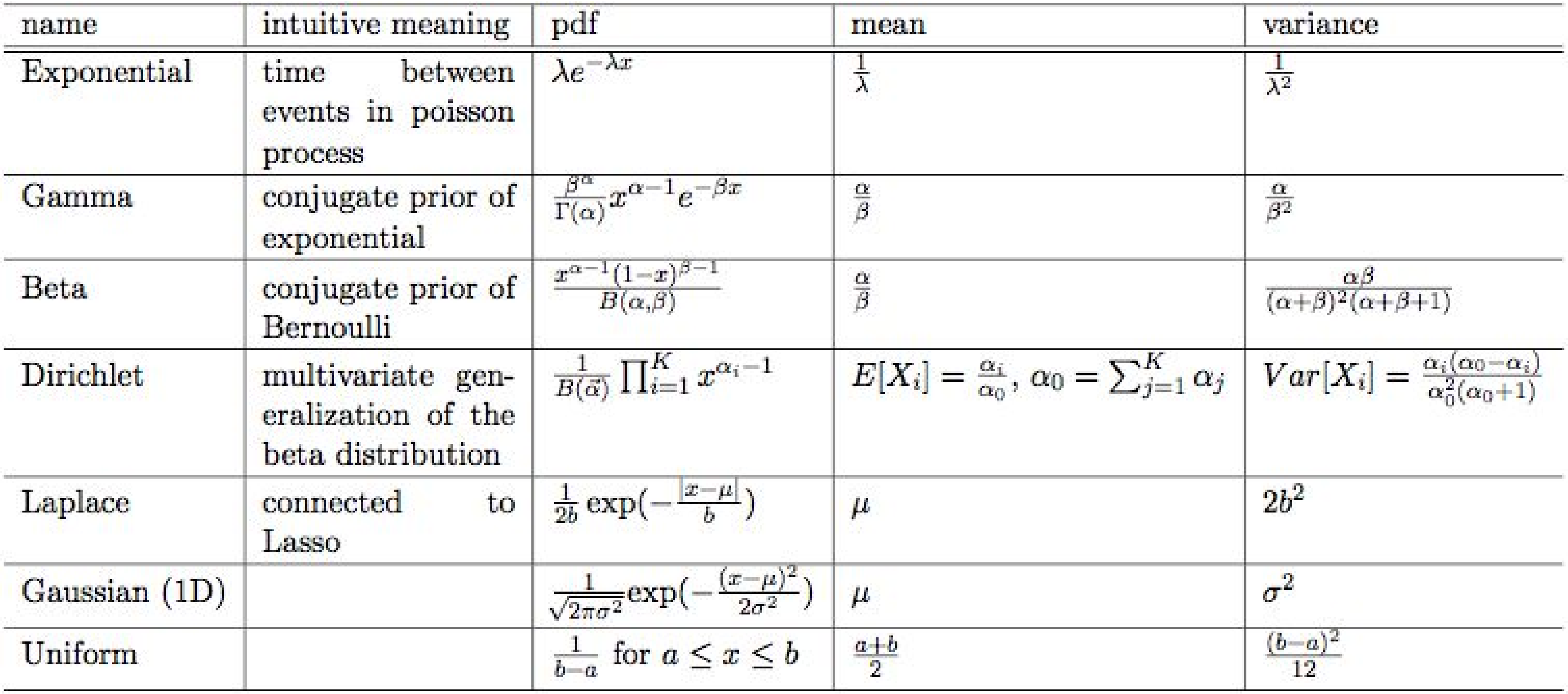
- Probability chain rule: $P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2) \ldots P(X_n \mid X_1, X_2, \ldots, X_{n-1})$
- Linearity of expectation: $E[aX + bY] = aE[X] + bE[Y]$
- Two forms of variance: $Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$
- $Cov[X, Y] = E[(X - E[X])(Y - E[Y])]$
- $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$
- When X, Y are independent, $Var[X + Y] = Var[X] + Var[Y]$, $E[XY] = E[X] \cdot E[Y]$, not true if they are not independent
- X, Y Independent $\implies Cov(X, Y) = 0$ (thats why covar matrix is diagonal), not the other way around
- Iterated expectation (total expectation): $E[Y] = E[E[Y \mid X]]$

Common distributions - discrete



name	intuitive meaning	pmf	mean	variance
bernoulli	a coin flip	$p^{\mathbb{I}(x=1)}(1-p)^{\mathbb{I}(x=0)}$	p	$p(1-p)$
binomial	k success in n flips	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
geometric	first success on the k-th flip	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$
categorical	a die	$p_1^{\mathbb{I}(x=1)}p_2^{\mathbb{I}(x=2)}\ldots p_k^{\mathbb{I}(x=k)}$		
multinomial	a sequence of die role	$\frac{n!}{x_1!x_2!\ldots x_k!p_1^{x_1}}p_2^{x_2}\ldots p_k^{x_k}$	$E[x_i]=np_i$	$Var[x_i]=np_i(1-p_i)$
uniform	a fair die	$\frac{1}{n}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
poisson	number of patient in fixed time t	$\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ

Common distributions - continuous



name	intuitive meaning	pdf	mean	variance
Exponential	time between events in poisson process	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	conjugate prior of exponential	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Beta	conjugate prior of Bernoulli	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$\frac{\alpha}{\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Dirichlet	multivariate generalization of the beta distribution	$\frac{1}{B(\vec{\alpha})} \prod_{i=1}^{K} x^{\alpha_i - 1}$	$E[X_i] = \frac{\alpha_i}{\alpha_0}, \alpha_0 = \sum_{j=1}^{K} \alpha_j$	$Var[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0+1)}$
Laplace	connected to Lasso	$\frac{1}{2b} \exp(-\frac{\lvert x-\mu \rvert}{b})$	μ	$2b^2$
Gaussian (1D)		$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$	μ	σ^2
Uniform		$\frac{1}{b-a}$ for $a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

Q1: Expectation

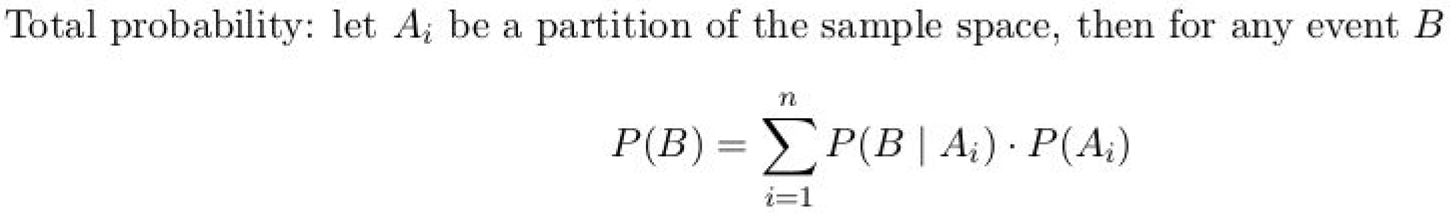
You are trapped in a dark cave with three indistinguishable exits on the walls. One of the exits takes you 3 hours to travel and takes you outside. One of the other exits takes 1 hour to travel and the other takes 2 hours, but both drop you back in the original cave. You have no way of telling which exits you have attempted. What is the expected time it takes for you to get outside?

Q1: Expectation

Let the random variable X be the time it takes for you to get outside. So, by the description of the problem, $E(X) = 1/3 * (3) + 1/3 (1 +E(X)) + 1/3 (2 +E(X))$. Solving this equation leads to the solution, $E(X) = 6$.

Q2: Total probability theorem

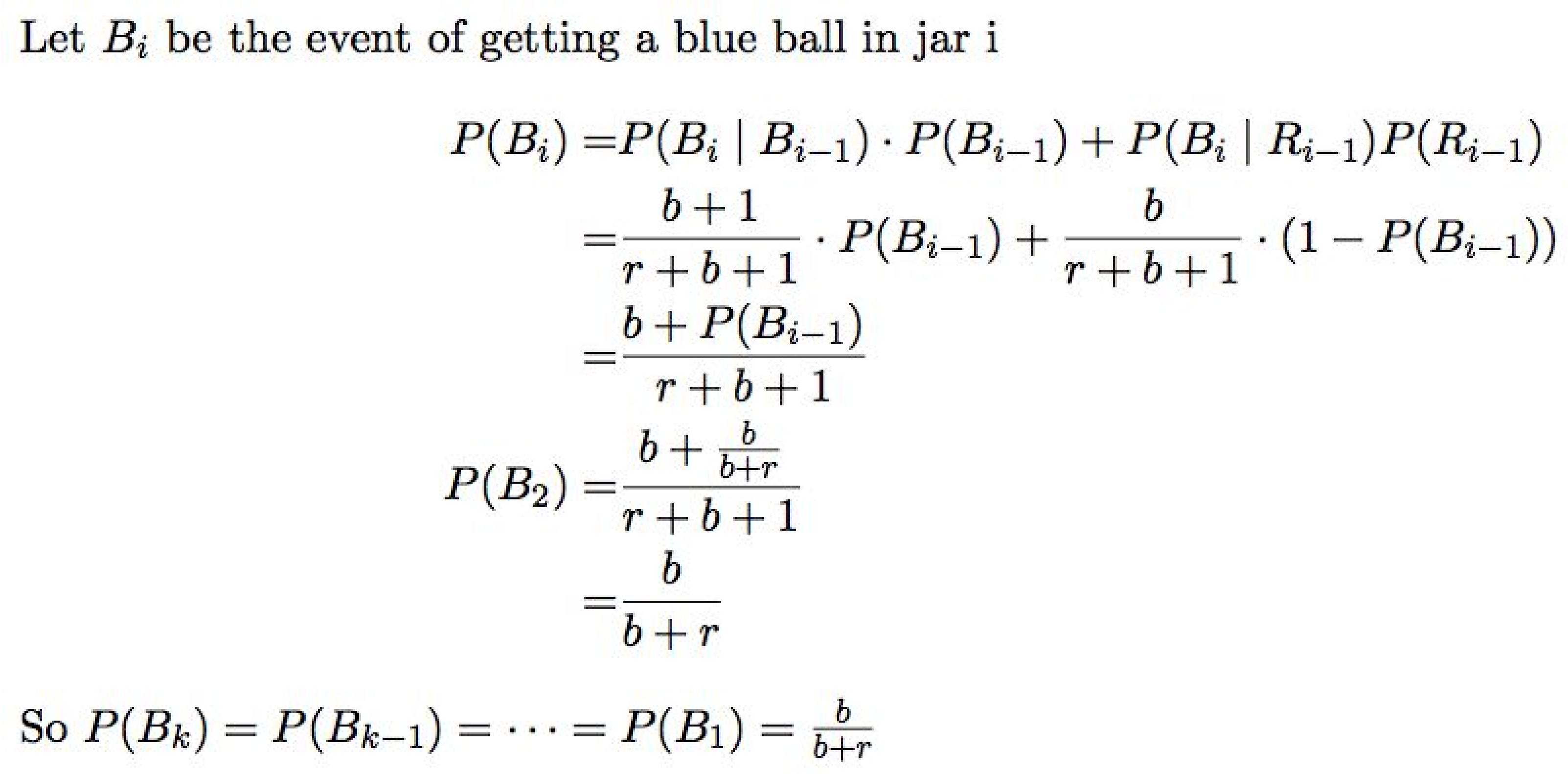
There are k jars, each containing r red balls and b blue balls. Randomly select a ball from jar 1 and transfer it to jar 2, then randomly select a ball from jar 2 and transfer to jar 3, ..., then randomly select a ball from jar (k - 1) and transfer to jar k. What's the probability that the last ball is blue?



Total probability: let A_i be a partition of the sample space, then for any event B

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)$$

Q2: Total probability theorem



Let B_i be the event of getting a blue ball in jar i

$$\begin{aligned}
P(B_i) =& P(B_i \mid B_{i-1}) \cdot P(B_{i-1}) + P(B_i \mid R_{i-1})P(R_{i-1}) \\
=& \frac{b+1}{r+b+1} \cdot P(B_{i-1}) + \frac{b}{r+b+1} \cdot (1 - P(B_{i-1})) \\
=& \frac{b + P(B_{i-1})}{r+b+1} \\
P(B_2) =& \frac{b + \frac{b}{b+r}}{r+b+1} \\
=& \frac{b}{b+r}
\end{aligned}$$

So $P(B_k) = P(B_{k-1}) = \cdots = P(B_1) = \frac{b}{b+r}$

MLE & MAP

Frequentist v/s Bayesian Statistics

Frequentist	Bayesian
An event's probability = Limit of its relative frequency in a large number of trials.	**An event's probability (posterior)** is a consequence of**:** - A **Prior** probability, and - A **Likelihood** Function derived from a statistical model for the observed data.
Maximum Likelihood Estimate (MLE)	Maximum a posteriori (MAP)

Maximum Likelihood Estimate

- We have some data ‘*D*’
- Which parameter / set of parameters make(s) *D* most probable

$$\theta^{MLE} = argmax_{\theta} P(D|\theta) = argmax_{\theta} \ln(P(D|\theta))$$

Problems:

- Bias due to undersampling
- 0-product due to undersampling

Maximum a posteriori

- We should choose the value of θ that is most probable, **given the observed data ‘*D*’** and our prior assumptions summarized by ***P(θ)***

$$\theta^{MAP} = argmax_{\theta} P(\theta|D) = argmax_{\theta} \frac{P(D|\theta)P(\theta)}{P(D)}$$

$$= argmax P(D|\theta)P(\theta)$$

Q1 - MLE for a Multinomial distribution

- Multinomial distribution : Generalized Binomial distribution
- It models the probability of counts for rolling a **K-sided die** *N* times

$$\theta^{\rightarrow} = \{\theta_1, \theta_2, ...\theta_K\}$$

$$\sum_{i=1}^{K} \theta_i = 1$$

Let $\boldsymbol{N}_i$ be the number of times face ***i*** of the die appeared and **N** be the total number of rolls. **What's the MAP estimate of the vector of parameters** ?

$$\theta^{\rightarrow MLE} = argmax P(D|\theta^{\rightarrow}) = argmax \prod_{i=1}^{K} \theta_i^{N_i}$$

$$= argmax \ln \prod_{i=1}^{K} \theta_i^{N_i} \qquad = l(\theta^{\rightarrow})$$

$$= argmax \sum_{i=1}^{K} N_i \ln \theta_i$$

Finding the MLE by setting the derivative to 0

$$\frac{\partial l}{\partial \theta_j} = \frac{\partial(\sum_{i=1}^{K} N_i \ln \theta_i)}{\partial \theta_j}$$

$$= \frac{\partial N_j \ln \theta_j}{\partial \theta_j}$$

$$= \frac{N_j}{\theta_j} \qquad = 0$$

$$\Rightarrow \theta_j = \infty$$

What happened ?
Did we mess up basic high-school calculus ?

Nah. We did not constrain the optimization problem !

- There are **2** ways to constrain the values of θ to ensure they fall between 0 and 1:
- Any ideas ?

1. Constraint : $\theta_K = 1 - \sum_{i=1}^{K-1} \theta_i$

$$\theta^{\rightarrow MLE} = argmax P(D|\theta^{\rightarrow}) = argmax \prod_{i=1}^{K} \theta_i^{N_i}$$

$$= argmax \left(\prod_{i=1}^{K-1} \theta_i^{N_i} \right) \theta_K^{N_K}$$

$$= argmax \left(\prod_{i=1}^{K-1} \theta_i^{N_i} \right) \left(1 - \sum_{i=1}^{K-1} \theta_i \right)^{N_K}$$

$$= argmax \ln \left[\left(\prod_{i=1}^{K-1} \theta_i^{N_i} \right) \left(1 - \sum_{i=1}^{K-1} \theta_i \quad \right)^{N_K} \right]$$

$$= argmax \left[\sum_{i=1}^{K-1} N_i \ln \theta_i + N_K \ln \left(1 - \sum_{i=1}^{K-1} \theta_i \quad \right) \right]$$

$$\frac{\partial l}{\partial \theta_j} = \frac{N_j}{\theta_j} + \frac{N_K}{1 - \sum_{i=1}^{K-1} \theta_i} \{-1\} = 0$$

$$\Rightarrow \theta_j = \frac{N_j}{N_K} * \theta_K$$

$$\theta_1 + \theta_2 + ...\theta_K = 1$$

$$\left[\frac{N_1}{N_K} + \frac{N_2}{N_K} + ... + \frac{N_K}{N_K}\right]\theta_K = 1$$

$$\Rightarrow \theta_K = \frac{N_K}{\sum_{i=1}^{K} N_i}$$

2. Method of Lagrange Multipliers

- Another way to solve a constrained optimization problem
- You are not expected to know this method for now.

Q2: Find the MAP estimate

- Say we flip a coin (with probability of heads =), '*N*' times and we get '*H*' number of heads and '*T*' number of tails.
- Assume coin flips are i.i.d
- Find the MAP estimate of θ given that we impose a Beta prior to overcome undersampling bias.

$$\theta^{MAP} = argmax_\theta P(D|\theta)P(\theta)$$

$$= argmax_\theta \left(\theta^H(1-\theta)^T\right)\left(\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{Beta(\alpha,\beta)}\right)$$

$$= argmax_\theta \left(\theta^H(1-\theta)^T\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$= argmax_\theta \theta^{H+\alpha-1}(1-\theta)^{T+\beta-1}$$

Looks familiar ?

$$argmax_\theta \theta^{H+\alpha-1}(1-\theta)^{T+\beta-1}$$

$$argmax_{\theta}\theta^{H+\alpha-1}(1-\theta)^{T+\beta-1}$$

- Same as the **MLE** estimate of probability of getting heads (θ)
- So what’s the closed-form answer ?

$$\theta^{MAP} = \frac{H + \alpha - 1}{H + \alpha - 1 + T + \beta - 1}$$

- You can think of α - 1 as ‘imaginary number of heads’ and **β**-1 as imaginary number of tails that form a part of your **prior belief** about what the distribution of heads and tails should be.

Naive Bayes

Q1: Counting the # of parameters

Consider a naive Bayes classifier with 3 boolean input variables, X1, X2 and X3, and one boolean output, Y .

- How many parameters must be estimated to train such a Naive Bayes classifier? (you need not list them unless you wish to, just give the total)
- How many parameters would have to be estimated to learn the above classifier if we **do not** make the Naive Bayes conditional independence assumption?

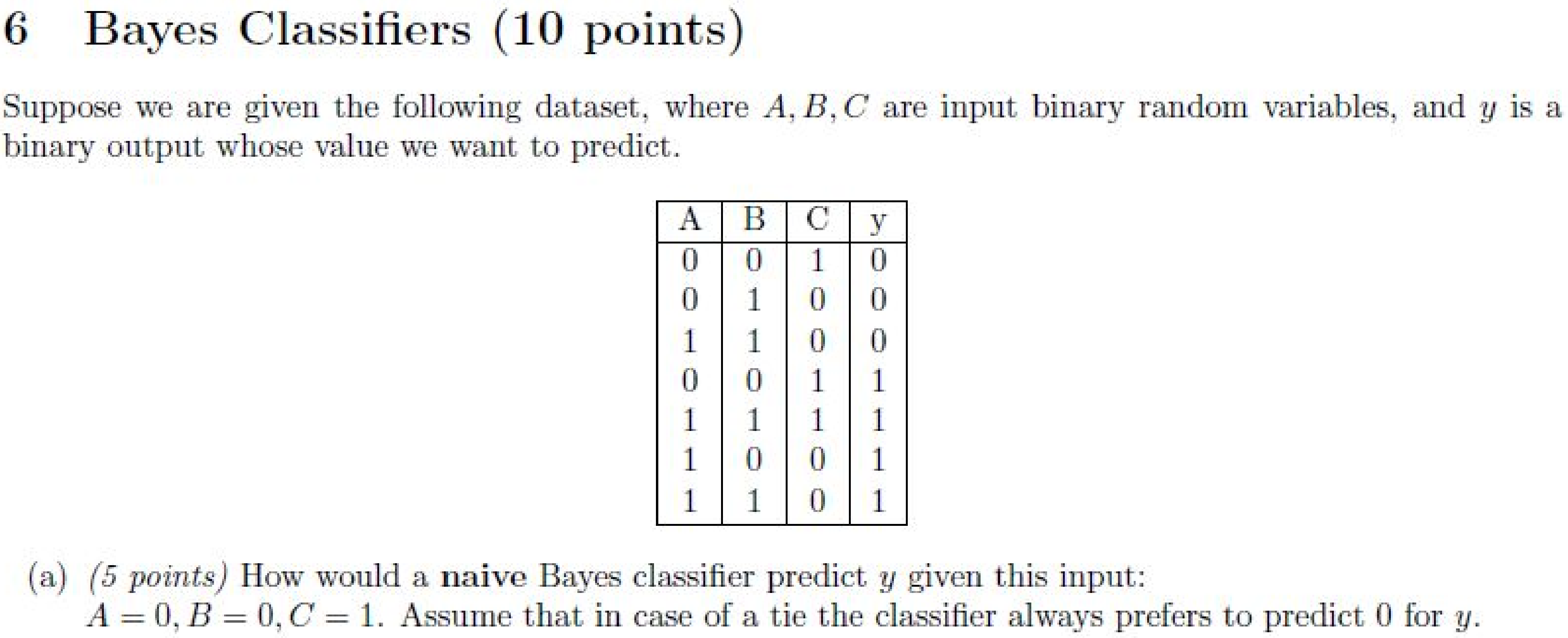
Q1: Counting the # of parameters

- Parameters needed for the **Naive Bayes** classifier:
  - P(Y=1)
  - P(X1 = 1ly = 0)
  - P(X2 = 1ly = 0)
  - P(X3 = 1ly = 0)
  - P(X1 = 1ly = 1)
  - P(X2 = 1ly = 1)
  - P(X3 = 1ly = 1).
- Other probabilities can be obtained with the constraint that the probabilities sum up to 1. So we need to estimate 7 parameters.

Q1: Counting the # of parameters

- Parameters needed **without the conditional independence assumption**:
  - We still need to estimate P(Y=1)
  - For Y=1, we need to know all the enumerations of (X1,X2,X3), i.e., 2^3 of possible (X1,X2,X3). Consider the constraint that the probabilities sum up to 1, we need to estimate $2^3 - 1 = 7$ parameters for Y=1
  - Similarly we need $2^3 - 1$ parameters for $Y = 0$
- Therefore the total number of parameters is $1 + 2(2^3 - 1) = 15$.

Q1: Bayes’ Decision Rule



6 Bayes Classifiers (10 points)

Suppose we are given the following dataset, where A, B, C are input binary random variables, and y is a binary output whose value we want to predict.

A	B	C	y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

(a) *(5 points)* How would a **naive** Bayes classifier predict y given this input: $A = 0, B = 0, C = 1$. Assume that in case of a tie the classifier always prefers to predict 0 for y.

Q2: Bayes' Decision Rule

Let $D = (A=0, B=0, C=1)$
To assign a label *y* to D, we have to find out which is greater: **$P(y=0|D)$ or $P(y=1|D)$**

From Bayes' Rule **$P(y=i|D) \propto P(D|y=i) * P(y = i)$**

From the *Naive* in Naive Bayes:

$$P(y = 0 \mid D) \propto P(A=0|y=0) * P(B=0|y=0) * P(C=1|y=0) * P(y = 0)$$

AND

$$P(y = 1 \mid D) \propto P(A=0|y=1) * P(B=0|y=1) * P(C=1|y=1) * P(y = 1)$$

Step 1: Training

1.1 Calculating priors

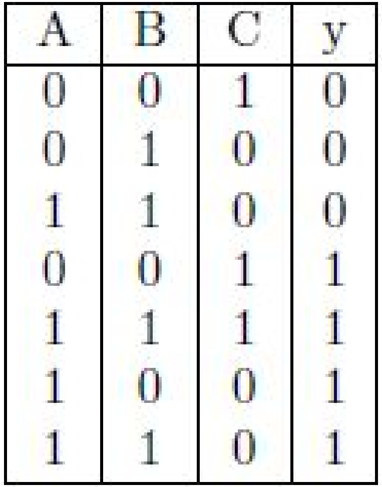
$P(y=1) = 4/7$

$P(y=0) = 1 - P(y=1)$

2.2 Estimating $P(X=X_i|y=y_i)$

	y = 0	y = 1
A= 0	2/3	1/4
B = 0	1/3	1/2
C =0	2/3	1/2

$1/4 \longrightarrow \boxed{P(A=0|y=1)}$



A	B	C	y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

Step 2: Predicting

$P(y = 0 \mid D) \propto P(A=0|y=0) * P(B=0|y=0) * P(C=1|y=0) * P(y = 0) = 0.0317$

$P(y = 1 \mid D) \propto P(A=0|y=1) * P(B=0|y=1) * P(C=1|y=1) * P(y = 1) = 0.0357$

Therefore predicted label = **1**

Another way to do this is log-sum