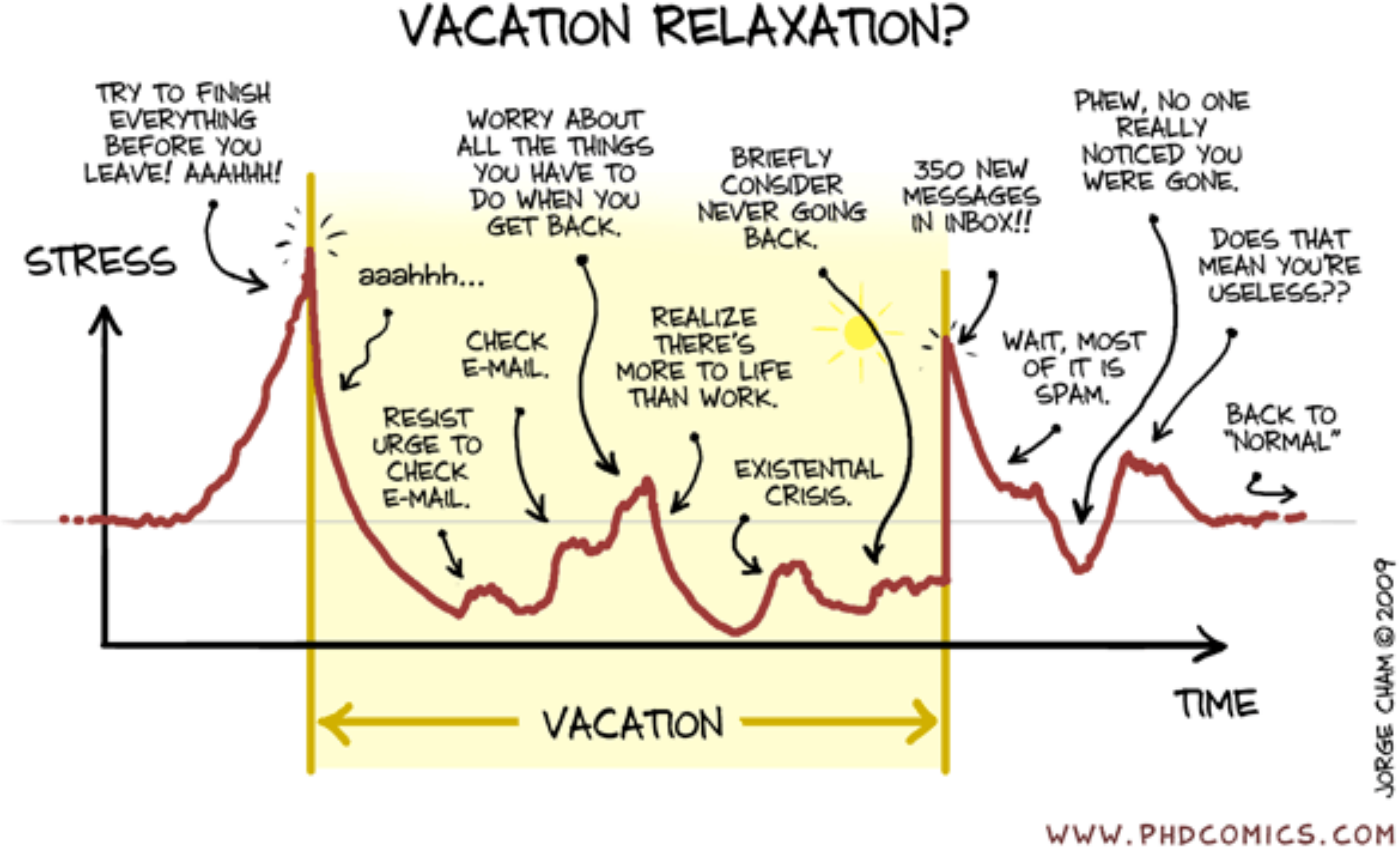
COMPUTER SCIENCE

INDIANA UNIVERSITY

School of Informatics and Computing
Bloomington

Parameter estimation
and
Intro to prediction problems


VACATION RELAXATION?
TRY TO FINISH EVERYTHING BEFORE YOU LEAVE! AAAHHH!
STRESS
aaahhh...
CHECK E-MAIL.
RESIST URGE TO CHECK E-MAIL.
WORRY ABOUT ALL THE THINGS YOU HAVE TO DO WHEN YOU GET BACK.
REALIZE THERE'S MORE TO LIFE THAN WORK.
BRIEFLY CONSIDER NEVER GOING BACK.
EXISTENTIAL CRISIS.
350 NEW MESSAGES IN INBOX!!
WAIT, MOST OF IT IS SPAM.
PHEW, NO ONE REALLY NOTICED YOU WERE GONE.
DOES THAT MEAN YOU'RE USELESS??
BACK TO "NORMAL"
VACATION
TIME
JORGE CHAM © 2009
WWW.PHDCOMICS.COM

Reminders

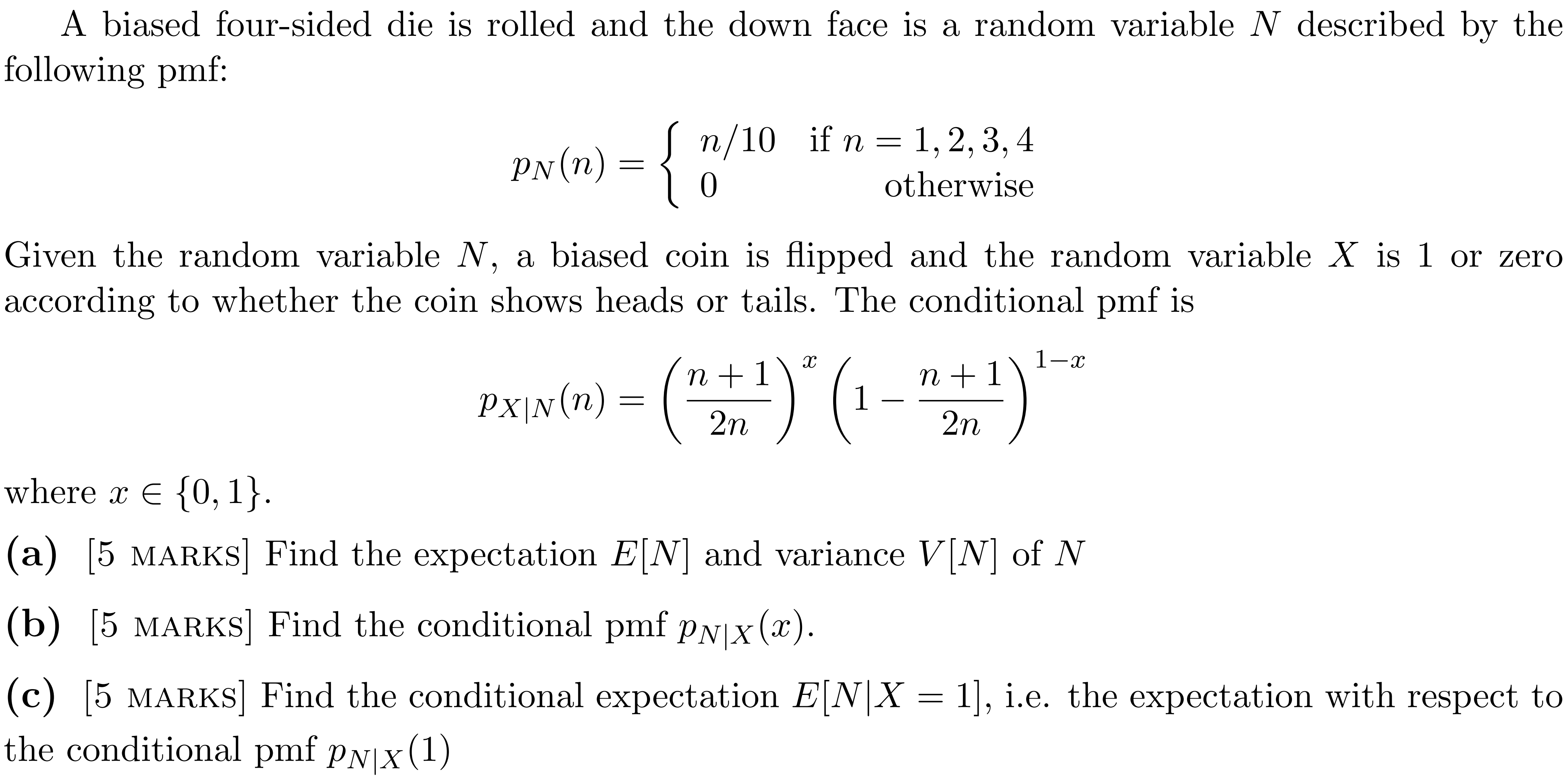
- My office hours are today from 2-4 p.m.
  - If anyone is feeling behind with probabilities, I can have a tutorial
- Thought questions #1 due today by midnight
  - you do not need to provide answers
  - I'll compile an anonymous list, so give the okay in your submission
- Assignment #1 is due next week on Wednesday
- Next readings: Chapters 3 and 4
  - Chapter 3: Introduction to machine learning
  - Chapter 4: Linear regression

Probability exercises

- Go through some exercise questions in the free textbook:
  - Bayesian Reasoning and Machine Learning, Barber

- Go through exercises in other recommended readings

- Example practice problem not included in homework



A biased four-sided die is rolled and the down face is a random variable N described by the following pmf:

$$p_N(n) = \begin{cases} n/10 & \text{if } n = 1,2,3,4 \\ 0 & \text{otherwise} \end{cases}$$

Given the random variable N, a biased coin is flipped and the random variable X is 1 or zero according to whether the coin shows heads or tails. The conditional pmf is

$$p_{X|N}(n) = \left(\frac{n+1}{2n}\right)^x \left(1 - \frac{n+1}{2n}\right)^{1-x}$$

where $x \in \{0,1\}$.

(a) [5 MARKS] Find the expectation $E[N]$ and variance $V[N]$ of N

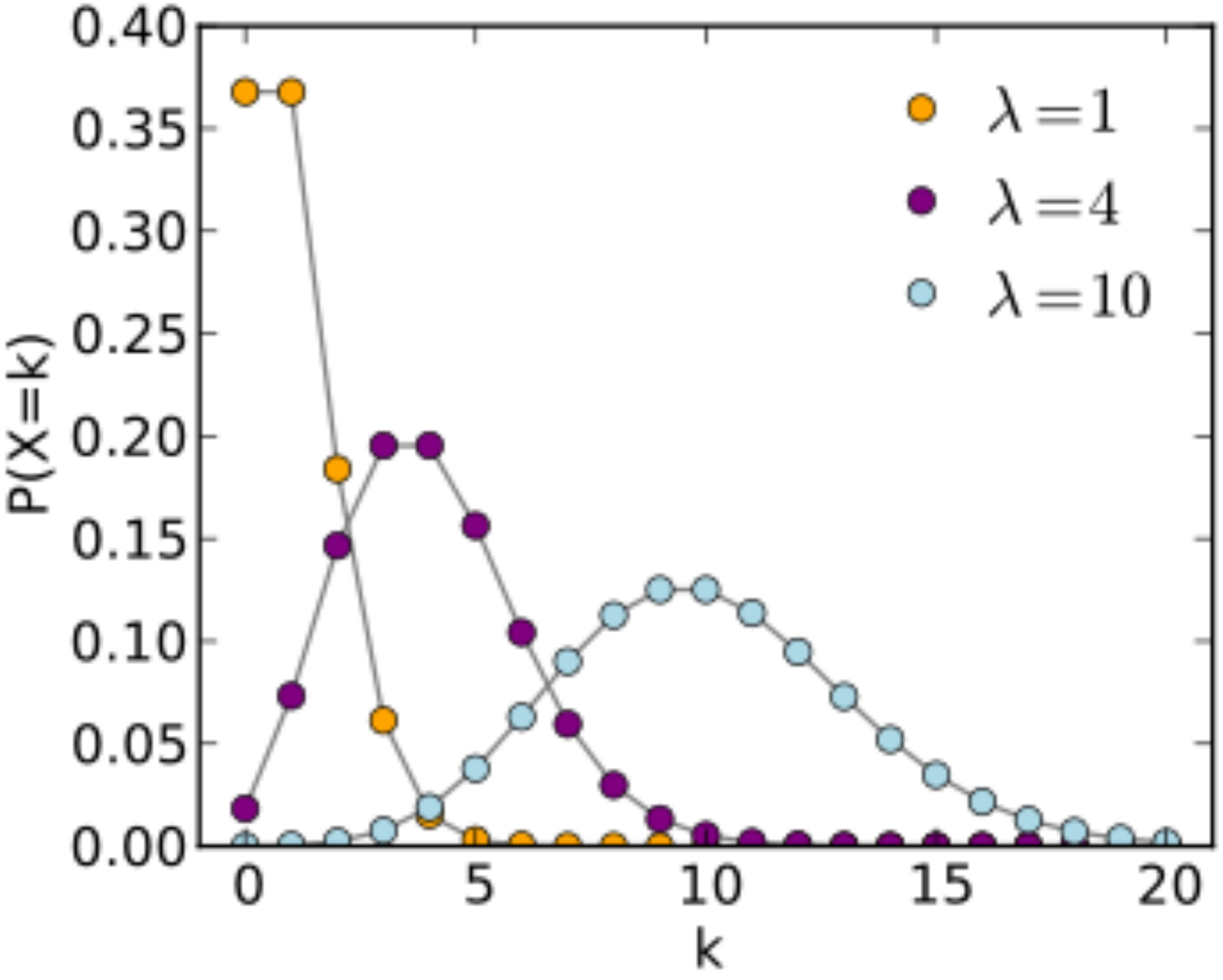
(b) [5 MARKS] Find the conditional pmf $p_{N|X}(x)$.

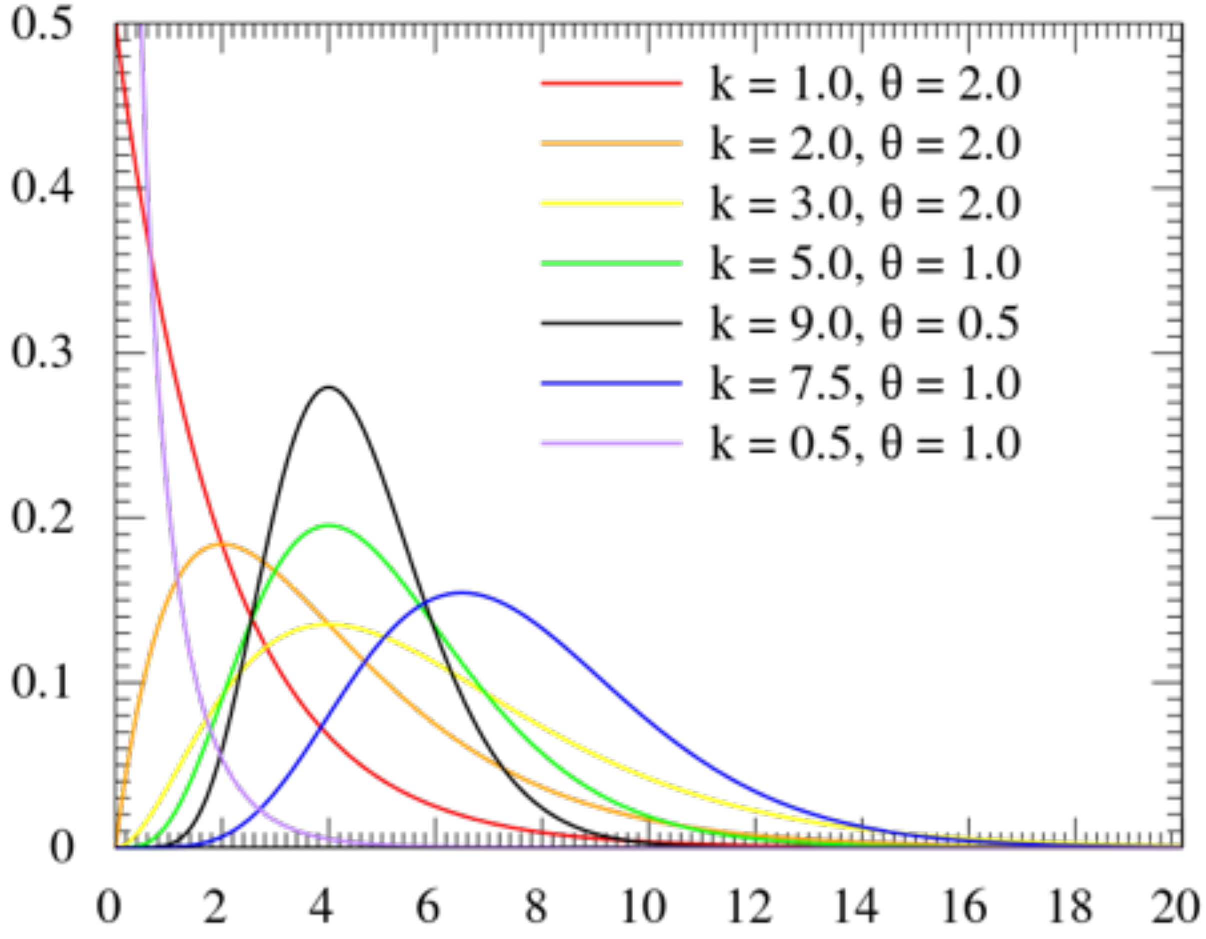
(c) [5 MARKS] Find the conditional expectation $E[N|X=1]$, i.e. the expectation with respect to the conditional pmf $p_{N|X}(1)$

Questions

- How did we select the parameters for the gamma prior on the Poisson parameters lambda for Example 9? Could these parameters instead be optimized?
  - this is an example of a good thought question
- How is hypothesis testing related to maximum likelihood?
  - look at likelihood ratio tests for one connection
- Review of interpretation and computing statistics like the mean and covariance —> we'll talk about this more today

Whiteboard: Example 9, 10





Poisson

Gamma

Whiteboard

- ML and MAP in limit of infinite samples
- Relationship of KL-divergence and ML

Prediction problem statement

- Data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$

$\mathbf{x}_i \in \mathcal{X}$ is the i-th object and $y_i \in \mathcal{Y}$ is the corresponding target designation

We usually assume that $\mathcal{X} = \mathbb{R}^d$, in which case $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$

- Each xi a data point or sample
- Each dimension of xi is called a feature or attribute
- Underlying assumption: features easy/easier to obtain and targets are difficult to observe or collect

Types of predictions

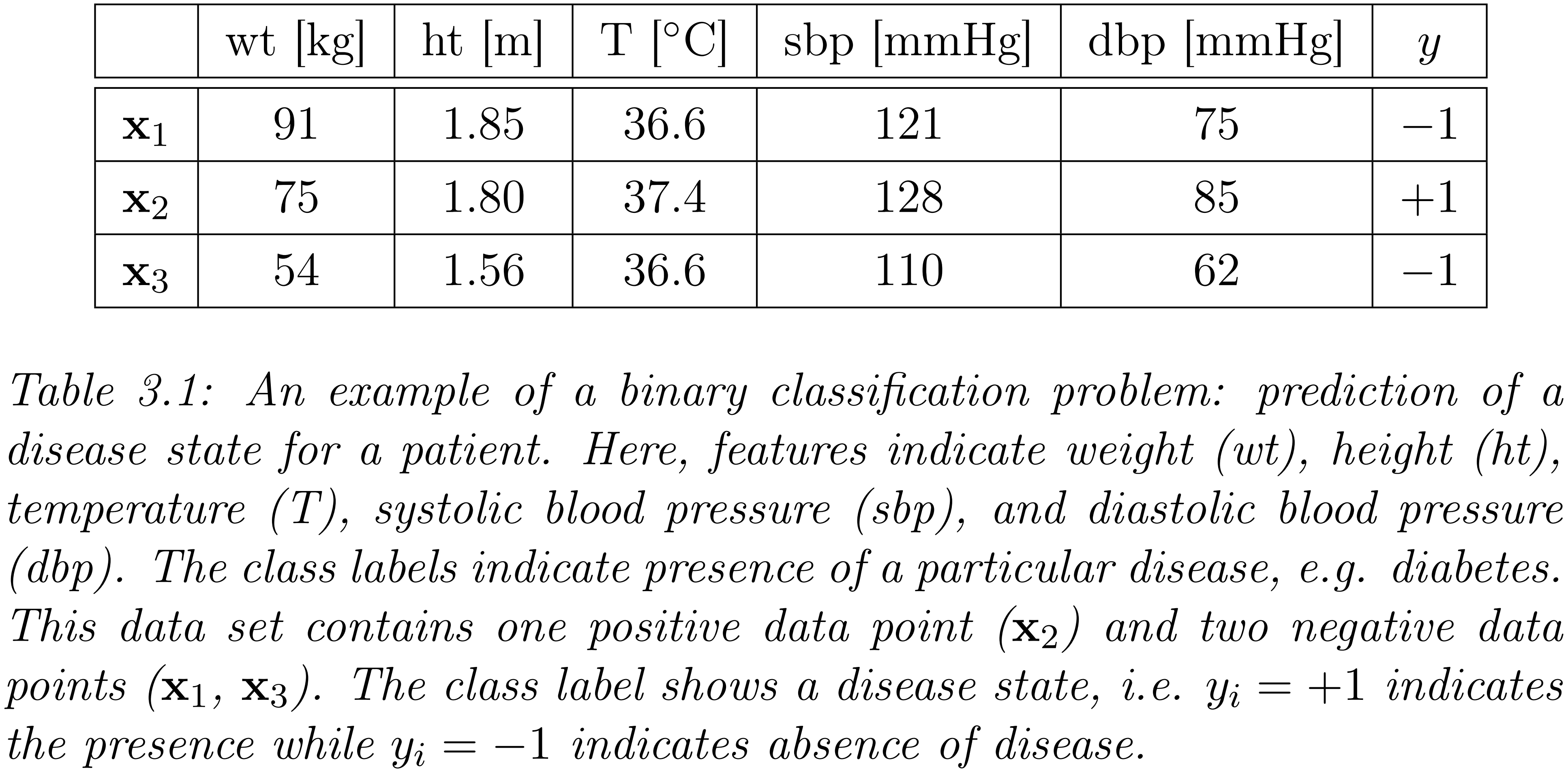
- The target could be anything; convenient to separate into different types (even though can have related approaches)

- Generally two main types discussed
  - classification, e.g. $\mathcal{Y} = \{\text{sports},\ \text{medicine},\ \text{travel}, \ldots\}.$
  - regression, e.g. $\mathcal{Y} = \mathbb{R}.$

- Structured output often a type of classification problem

- Can have structured output for regression as well

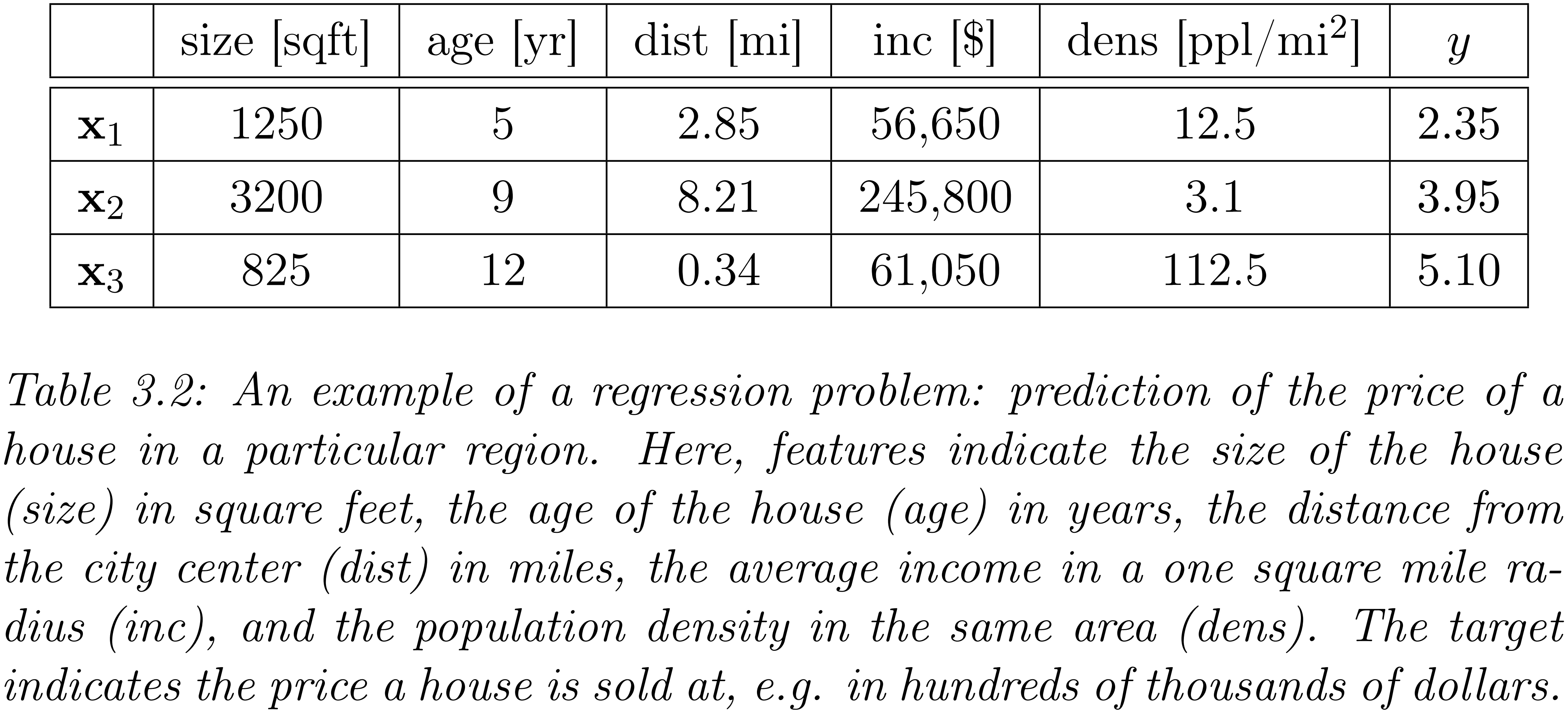
Example: binary classification



	wt [kg]	ht [m]	T [°C]	sbp [mmHg]	dbp [mmHg]	y
$\mathbf{x}_1$	91	1.85	36.6	121	75	-1
$\mathbf{x}_2$	75	1.80	37.4	128	85	$+1$
$\mathbf{x}_3$	54	1.56	36.6	110	62	-1

Table 3.1: An example of a binary classification problem: prediction of a disease state for a patient. Here, features indicate weight (wt), height (ht), temperature (T), systolic blood pressure (sbp), and diastolic blood pressure (dbp). The class labels indicate presence of a particular disease, e.g. diabetes. This data set contains one positive data point ($\mathbf{x}_2$) and two negative data points ($\mathbf{x}_1$, $\mathbf{x}_3$). The class label shows a disease state, i.e. $y_i = +1$ indicates the presence while $y_i = -1$ indicates absence of disease.

Example: regression



	size [sqft]	age [yr]	dist [mi]	inc [$]	dens [ppl/mi^2]	y
$\mathbf{x}_1$	1250	5	2.85	56,650	12.5	2.35
$\mathbf{x}_2$	3200	9	8.21	245,800	3.1	3.95
$\mathbf{x}_3$	825	12	0.34	61,050	112.5	5.10

Table 3.2: An example of a regression problem: prediction of the price of a house in a particular region. Here, features indicate the size of the house (size) in square feet, the age of the house (age) in years, the distance from the city center (dist) in miles, the average income in a one square mile radius (inc), and the population density in the same area (dens). The target indicates the price a house is sold at, e.g. in hundreds of thousands of dollars.

Other examples

- In general, as with much of machine learning, there are no strict rules: the way you specify a problem can be an art
  - consequences for learning can be strongly impacted by specification
- E.g: multi-label classification problem $\mathcal{Y} = \{\text{sports}, \ \text{medicine}, \ \text{travel}, \ldots\}$.
  - could specify as high-dimensional multi-class classification

$$\mathcal{Y} = \mathcal{P}(\{\text{sports}, \ \text{medicine}, \ \text{travel}, \ldots\})$$

  - could specify as a regression problem
  - could try to learn hierarchical tree structure
- A key part of ML is not only knowing algorithms, but also how to formalize (prediction) problems

Vector and matrix notation

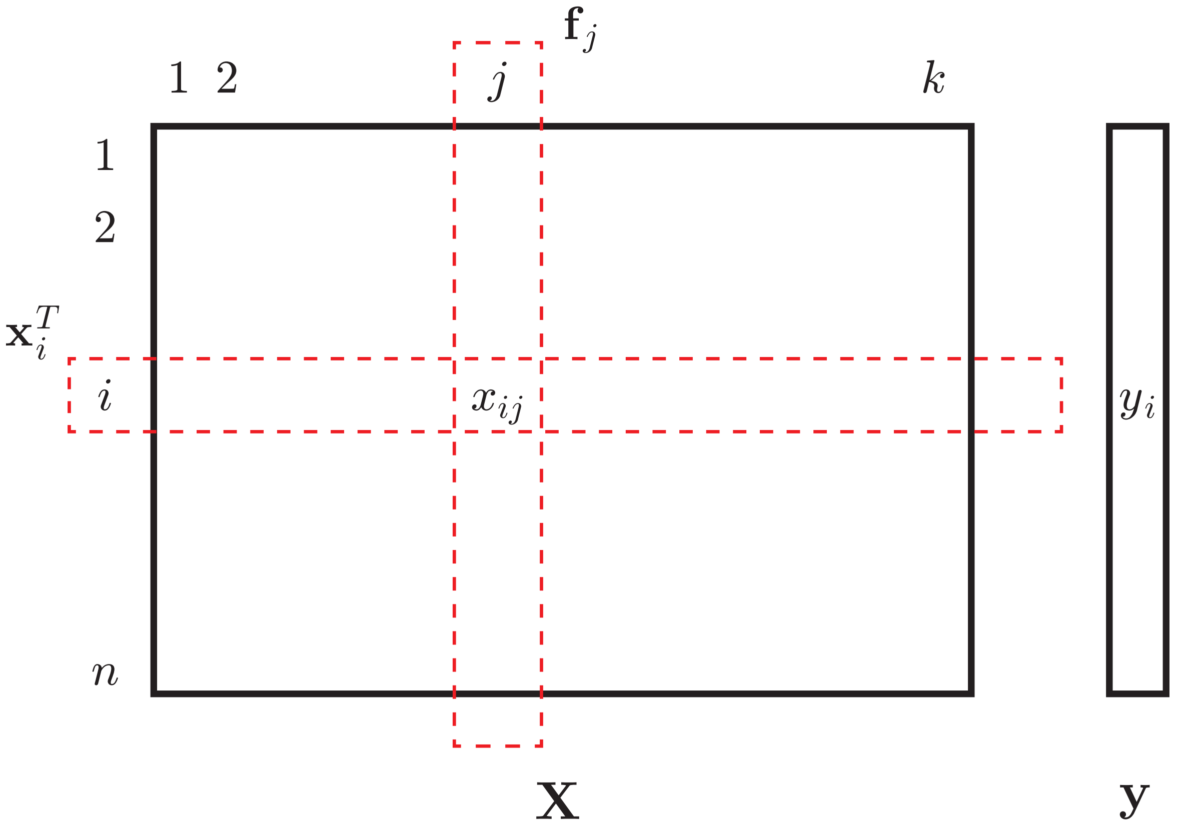


Figure 3.1: Data set representation and notation. $\mathbf{x}$ *is an* n*-by-*k *matrix representing features and data points, whereas* y *is an* n*-by-*1 *vector of targets.*

Thought exercise: specifying prediction problems

- Imagine someone has given you a database of samples
  - e.g., Netflix data of people’s movie rankings
- What do you need to consider before starting to learn a predictor? (without knowing much about algorithms yet)
  - How do I know if my predictions are successful? What is my measure?
  - How many samples are there? Is it a large database?
  - Is efficiency important?
  - What simple algorithms can I try first?
  - Is the data useful? Could it be significantly improved with different and/or more data collection?
  - How should I represent my prediction problem?

Optimal prediction

- We want to learn a prediction function
- We will need to define cost/error of a prediction
    - Cost function $c : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty)$
    - For true target y, get cost $c(\hat{y}, y)$
- We will see that modeling p(y | x) is useful for this task
- Want to find predictor that minimizes the expected cost
    - could choose other metrics, such as minimize number of costs that are very large or minimize the maximum cost

Whiteboard

- Expected cost
- Bayes optimal models
- Next topic: linear regression