Orsolya Putz, Zoltán Varjú

Text Analytics with Python for Cognitive and Social Scientists

Contents

List of Figures

List of Tables

Part I

Basic Concepts

1

Text Analytics

CONTENTS

A component part for an electronic item is manufactured at one of three different factories, and then delivered to the main assembly line.Of the total number supplied, factory A supplies 50%, factory B 30%, and factory C 20%. Of the components manufactured at factory A, 1% are faulty and the corresponding proportions for factories B and C are 4% and 2% respectively. A component is picked at random from the assembly line. What is the probability that it is faulty?

1.1 Introduction

The term reliability usually refers to the probability that a component or system will operate satisfactorily either at any particular instant at which it is required or for a certain length of time. Fundamental to quantifying reliability s a knowledge of how to define, assess and combine probabilities [**?**]. This may hinge on identifying the form of the variability which is nherent n most processes. If all components had a fixed known lifetime there would be no need to model reliability.

1. A component part for an electronic item is manufactured at one of three different factories.
2. A component part x for an electronic item is manufactured at one of three different factories.
   (a) A component part for an electronic item is manufactured at one of three different factories.
   (b) A component part x for an electronic item is manufactured at one of three different factories.
      i. A component part for an electronic item is manufactured at one of three different factories.

ii. A component part x for an electronic item is manufactured at one of three different factories.
iii. A component part for an electronic item is manufactured at one of three different factories.
iv. A component part 1, 2, 3, 4 for an electronic item is manufactured at one of three different factories.
v. A component part for enumerate list of an electronic item is manufactured at one of three different factories.

(c) A component part for an electronic item is manufactured at one of three different factories.
(d) A component part 1, 2, 3, 4 for an electronic item is manufactured at one of three different factories.
(e) A component part for enumerate list of an electronic item is manufactured at one of three different factories.

3. A component part for an electronic item is manufactured at one of three different factories.
4. A component part 1, 2, 3, 4 for an electronic item is manufactured at one of three different factories.
5. A component part for enumerate list of an electronic item is manufactured at one of three different factories.

1.1.1 A component part

A component part for an electronic item is manufactured at one of three different factories, and then delivered to the main assembly line.Of the total number supplied, factory A supplies 50%, factory B 30%, and factory C 20%. Of the components manufactured at factory A, 1% are faulty and the corresponding proportions for factories B and C are 4% and 2% respectively. A component is picked at random from the assembly line. What is the probability that it is faulty [**?**]? A component part for an electronic item is manufactured at one of three different factories, and then delivered to the main assembly line. Of the total number supplied, factory A supplies 50%, factory B 30%, and factory C 20%. Of the components manufactured at factory A, 1% are faulty and the corresponding proportions for factories B and C are 4% and 2% respectively. A component is picked at random from the assembly line. What is the probability that it is faulty? A component part for an electronic item is manufactured at one of three different factories, and then delivered to the main assembly line.Of the total number supplied, factory A supplies 50%, factory B 30%, and factory C 20%. Of the components manufactured at factory A, 1% are faulty and the corresponding proportions for factories B and C are 4% and 2% respectively. A component is picked at random from the assembly line. What is the probability that it is faulty?

TABLE 1.1
Now we are engaged (a_g^a) $\left(a_g^a\right)$ in a great civil war, testing whether that nation, or any nation so conceived.

Scene	Reg. fts.	Hor. fts.	Ver. fts.
Ball	19, 221	4, 598	3, 200
Pepsi[a]	46, 281	6, 898	5, 400
Keybrd[b]	27, 290	2, 968	3, 405
Pepsi	14, 796	9, 188	3, 209

"A Process is a structured, measured set of activities designed to produce a specific output for a particular customer or market—A process is thus a specific ordering of work activities across time and space, with a beginning, an end. and clearly defined inputs and outputs: a structure for action."

Thomas Davenport
Senior Adjutant to the Junior Marketing VP

MultiRelational k-Anonymity. Most works on k-anonymity focus on anonymizing a single data table; however, a real-life [**?**] database usually contains multiple relational tables. This has proposed a privacy model called *MultiR k-anonymity* to ensure k-anonymity on multiple relational tables. Their model assumes that a relational database contains a person-specific table PT and a set of tables $T_1, \cdots, T_n$, where PT contains a person identifier Pid and some sensitive attributes, and T_i, for $1 \leq i \leq n$, contains some foreign keys, some attributes in QID, and sensitive attributes. The general privacy notion is to ensure that for each record owner o contained in the join of all tables $PT \bowtie T_1 \bowtie \cdots \bowtie T_n$, there exists at least $k-1$ other record owners share the same QID with o. It is important to emphasize that the k-anonymization is applied at the *record owner* level, not at the *record* level in traditional k-anonymity. This idea is similar to (X, Y)-anonymity, where $X = QID$ and $Y = \{Pid\}$.

Most works on k-anonymity focus on anonymizing a single data table; however, a real-life [**?**] database usually contains multiple relational tables. This has proposed a privacy model called *MultiR k-anonymity* to ensure k-anonymity on multiple relational tables. Their model assumes that a relational database contains a person-specific table PT and a set of tables $T_1, \cdots, T_n$, where PT contains a person identifier Pid and some sensitive attributes, and T_i, for $1 \leq i \leq n$, contains some foreign keys, some attributes in QID, and sensitive attributes. The general privacy notion is to ensure that for each

record owner o contained in the join of all tables $PT \bowtie T_1 \bowtie \cdots \bowtie T_n$, there exists at least $k-1$ other record owners share the same QID with o. It is important to emphasize that the k-anonymization is applied at the *record owner* level, not at the *record* level in traditional k-anonymity. This idea is similar to (X, Y)-anonymity, where $X = QID$ and $Y = \{Pid\}$.

In most literature on PPDP, they [**?**] consider a more relaxed, yet more practical, notion of privacy protection by assuming limited attacker's background knowledge. Below, the term "victim" refers to the record owner being linked. We can broadly classify linking models to two families.

> A component part for an electronic item is [**?**] manufactured at one of three different factories, and then delivered to the main assembly line.Of the total number supplied, factory A supplies 50%, factory B 30%, and factory C 20%. Of the components manufactured at factory A, 1% are faulty and the corresponding proportions for factories B and C are 4% and 2% respectively.

One family considers a privacy threat occurs when an attacker is able to link a record owner to a record in a published data table, to a sensitive attribute in a published data table, or to the published data table itself. We call them *record linkage*, *attribute linkage*, and *table linkage*, respectively. In all types of linkages, we assume that the attacker knows the QID of the victim. In record and attribute linkages, we further assume that the attacker knows the presence of the victim's record in the released table, and seeks to identify the victim's record and/or sensitive information from the table [**?**]. In table linkage, the attack seeks to determine the present or absent of the victim's record in the released table. A data table is considered to privacy preserved if the table can effectively prevent the attacker from successfully performing these types of linkages on the table [**?**]. Sections **??**-**??** study this family of

TABLE 1.2
Now we are engaged (a_g^a) $\left(a_g^a\right)$ in a great civil war, testing whether that nation, or any nation so conceived.

Scene	Reg. fts.	Hor. fts.	Ver. fts.
Table Head			
Ball	19, 221	4, 598	3, 200
Pepsi	46, 281	6, 898	5, 400
Keybrd	27, 290	2, 968	3, 405
Pepsi	14, 796	9, 188	3, 209

privacy models.

$$\text{var}\widehat{\Delta} = \sum_{j=1}^{t} \sum_{k=j+1}^{t} \text{var}\,(\hat{\alpha}_j - \hat{\alpha}_k) = \sum_{j=1}^{t} \sum_{k=j+1}^{t} \sigma^2(1/n_j + 1/n_k). \tag{1.1}$$

An obvious measure of imbalance is just the difference in the number of times the two treatments are allocated

$$D_n = \mathcal{M}|n_A - n_B|. \tag{1.2}$$

For rules such as deterministic allocation, for which the expected value of this difference can be calculated, we obtain the population value $\mathcal{D}_n$.

Box Title Here

Another family aims at achieving the *uninformative principle*: The published table should provide the attacker with little additional information beyond the background knowledge. There should not be a large difference between the prior and posterior beliefs; otherwise, there is a privacy threat [**?**, **?**]. Many privacy models in this family are designed for statistical database and do not distinguish attributes in T into QID, but some of them could also thwart record, attribute, and table linkages. Section **??** studies this family of privacy models.

Let m be a prime number. With the addition and multiplication as defined above, Z_m is a field.

Theorem 1 *Let m be a prime number. With the addition and multiplication as defined above, Z_m is a field.*

Proof 1 *Most of the proof of this theorem is routine. It is clear that $0 \in Z_m$ and $1 \in Z_m$ are the zero element and identity element. If $a \in Z_m$ and $a \neq 0$, then $m - a$ is the additive inverse of a. If $a \in Z_m$ and $a \neq 0$, then the greatest common divisor of a and m is 1, and hence there exist integers s and t such that $sa + tm = 1$. Thus $sa = 1 - tm$ is congruent to 1 modulo m. Let s^* be the integer in Z_m congruent to s modulo m. Then we also have $s^*a \equiv 1$ mod m. Hence s^* is the multiplicative inverse of a modulo m. Verification of the rest of the field properties is now routine.*

1.2 Record Linkage Model

In the privacy attack of *record linkage*, some value *qid* on QID identifies a small number of records in the released table T, called a *group*. If the victim's

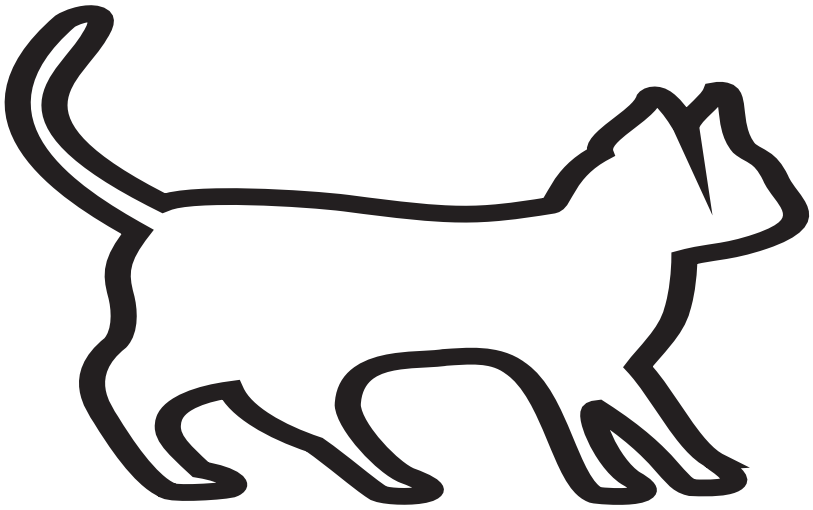

FIGURE 1.1
Figure caption goes here.

QID matches the value qid, the victim is vulnerable to being linked to the small number of records in the group [**?**]. In this case, the attacker faces only a small number of possibilities for the victim's record, and with the help of additional knowledge, there is a chance that the attacker could uniquely identify the victim's record from the group.

1.2.1 A component part

A component part for an electronic item is manufactured at one of three different factories, and then delivered to the main assembly line.Of the total number supplied, factory A supplies 50%, factory B 30%, and factory C 20%. Of the components manufactured at factory A, 1% are faulty and the corresponding proportions for factories B and C are 4% and 2% respectively. A component is picked at random from the assembly line. What is the probability that it is faulty?

1.2.1.1 H3 A component part

A component part for an electronic item is manufactured at one of three [**?**] different factories, and then delivered to the main assembly line.Of the total number supplied, factory A supplies 50%, factory B 30%, and factory C 20%. Of the components manufactured at factory A, 1% are faulty and the corresponding proportions for factories B and C are 4% and 2% respectively. A

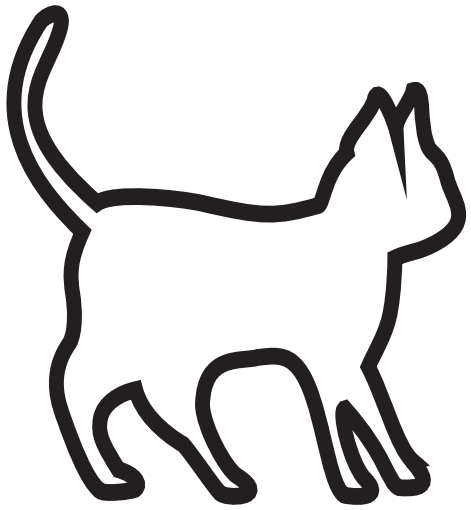

FIGURE 1.2
Figure caption goes here. Figure caption goes here. Figure caption goes here. Figure caption goes here. Figure caption goes here. Figure caption goes here.

component is picked at random from the assembly line. What is the probability that it is faulty?

A fundamental notion [**?**] is that of a subspace of F^n. Let V be a nonempty subset of F^n. Then V is a *subspace* of F^n provided V is closed under vector addition and scalar multiplication, that is,

(a) For all u and v in V, $u + v$ is also in V.

(b) For all u in V and c in F, cu is in V.

Let u be in the subspace V. Because $0u = 0$, it follows that the zero vector is in V. Similarly, $-u$ is in V for all u in V. A simple example of a subspace of F^n is the set of all vectors $(0, a_2, \ldots, a_n)$ with first coordinate equal to 0. The zero vector itself is a subspace.

Definition 1 Let $u^{(1)}, u^{(2)}, \ldots, u^{(m)}$ be vectors in F^n, and let $c_1, c_2, \ldots, c_m$ be scalars. Then the vector

$$c_1 u^{(1)} + c_2 u^{(2)} + \cdots + c_m u^{(m)}$$

is called a *linear combination* of $u^{(1)}, u^{(2)}, \ldots, u^{(m)}$. If V is a subspace of F^n, then V is closed under vector addition and scalar multiplication, and it follows easily by induction that a linear combination of vectors in V is also a vector in V. Thus *subspaces are closed under linear combinations*; in fact, this can be taken as the defining property of subspaces. The vectors

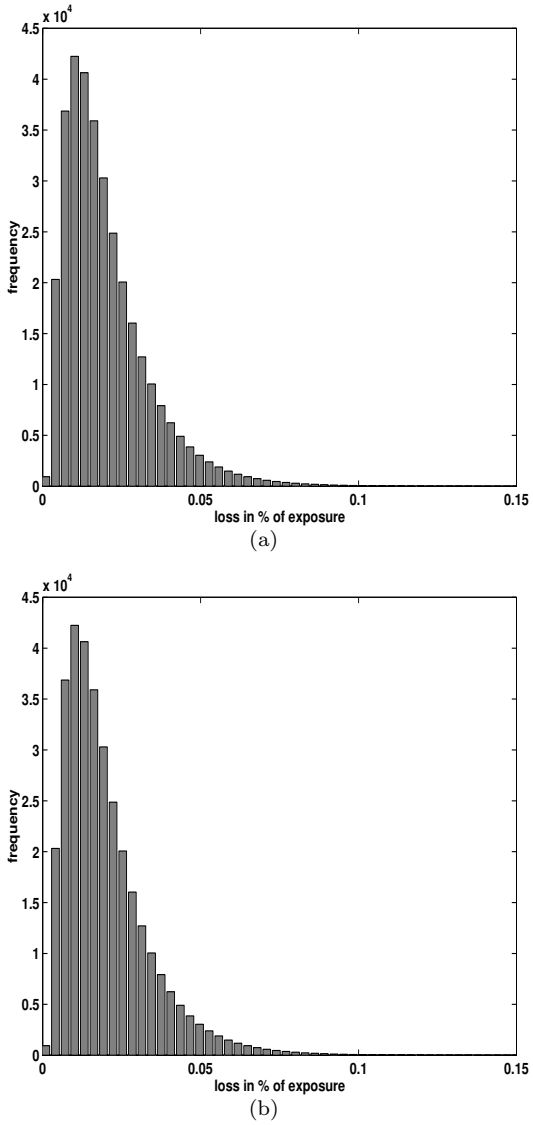

FIGURE 1.3
The bar charts depict the different risk contributions (top: 99% quantile, bottom: 99.9% quantile) of the business areas of a bank. The black bars are based on a Var/Covar approach, the white ones correspond to shortfall risk.

$u^{(1)}, u^{(2)}, \ldots, u^{(m)}$ *span* V (equivalently, form a *spanning set* of V) provided every vector in V is a linear combination of $u^{(1)}, u^{(2)}, \ldots, u^{(m)}$. The zero vector can be written as a linear combination of $u^{(1)}, u^{(2)}, \ldots, u^{(m)}$ with all scalars equal to 0; this is a *trivial linear combination*. The vectors $u^{(1)}, u^{(2)}, \ldots, u^{(m)}$ are *linearly dependent* provided there are scalars $c_1, c_2, \ldots, c_m$, not all of which are zero, such that

$$c_1 u^{(1)} + c_2 u^{(2)} + \cdots + c_m u^{(m)} = 0,$$

that is, the zero vector can be written as a *nontrivial linear combination* of $u^{(1)}, u^{(2)}, \ldots, u^{(m)}$. For example, the vectors $(1, 4), (3, -1)$, and $(3, 5)$ in $\Re^2$ are linearly dependent since

$$3(1, 4) + 1(3, -2) - 2(3, 5) = (0, 0).$$

Vectors are *linearly independent* provided they are not linearly dependent. The vectors $u^{(1)}, u^{(2)}, \ldots, u^{(m)}$ are a *basis* of V provided they are linearly independent and span V. By an *ordered basis* we mean a basis in which the vectors of the basis are listed in a specified order; to indicate that we have an ordered basis we write $(u^{(1)}, u^{(2)}, \ldots, u^{(m)})$. A spanning set S of V is a *minimal spanning set of* V provided that each set of vectors obtained from S by removing a vector is not a spanning set for V. A linearly independent set S of vectors of V is a *maximal linearly independent set of vectors of* V provided that for each vector w of V that is not in S, $S \cup \{w\}$ is linearly dependent (when this happens, w must be a linear combination of the vectors in S). □

In addition to matrix addition, subtraction, and multiplication, there is one additional operation that we define now. It's perhaps the simplest of them all. Let $A = [a_{ij}]$ be an m by n matrix and let c be a number [**?**]. Then the matrix $c \cdot A$, or simply cA, is the m by n matrix obtained by multiplying each entry of A by c:

$$cA = [ca_{ij}].$$

The matrix cA is called a *scalar multiple* of A.

Think About It...

Commonly thought of as the first modern computer, ENTAC was built in 1944. It took up more space than an 18-wheeler's tractor trailer and weighed more than 17 Chevrolet Camaros. It consumed 140,000 watts of electricity while executing up to 5,000 basic arithmetic operations per second. One of today's popular microprocessors, the 486, is built on a tiny piece of silicon about the size of a dime.

With the continual expansion of capabilities, computing power

will eventually exceed the capacity for human comprehension or human control.

The Information Revolution
Business Week

1.3 Glossary

360 Degree Review: Performance review that includes feedback from superiors, peers, subordinates, and clients.

Abnormal Variation: Changes in process performance that cannot be accounted for by typical day-to-day variation. Also referred to as nonrandom variation.

Acceptable Quality Level (AQL): The minimum number of parts that must comply with quality standards, usually stated as a percentage.

Activity: The tasks performed to change inputs into outputs.

Adaptable: An adaptable process is designed to maintain effectiveness and efficiency as requirements change. The process is deemed adaptable when there is agreement among suppliers, owners, and customers that the process will meet requirements throughout the strategic period.

Part II

Graph-based methods

Part III

Embeddings

Part IV

Advanced topics