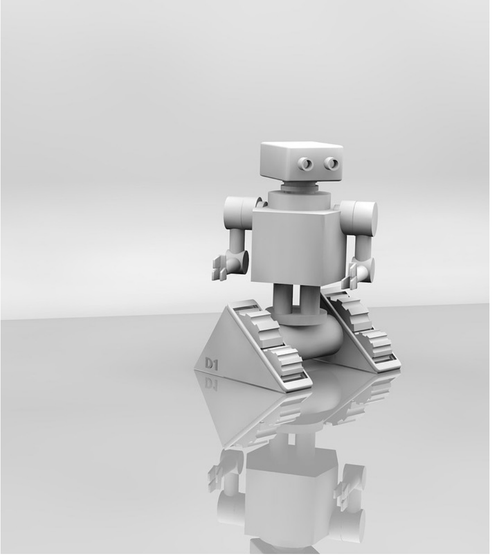


CPSC 481
Artificial Intelligence

Dr. Mira Kim
Mira.kim@fullerton.edu

What we will cover this week

- Reasoning with uncertainty
  - Probability
  - Conditional probability
  - Bayes rule
  - Naïve Bayes classifier

Applications with uncertainty

- Medical diagnosis
  - Not always an obvious cause/effect relationship between the symptoms presented by the patient and the causes of these symptoms
  - The same sets of symptoms from multiple causes
- Natural language understanding
  - Words, expressions, and metaphors are learned, but also change and evolve as they are used over time.
- Planning and scheduling
  - When an agent forms a plan (e.g., a vacation trip by car), often no deterministic sequence of operations is guaranteed to succeed
  - Car can car breaks down, the car ferry is cancelled, a hotel is fully booked even though a reservation was made
- Learning
  - An important component of many stochastic systems is that they have the ability to sample situations and learn over time

Uncertainty

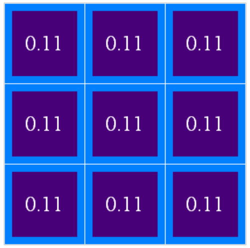
- General situation:

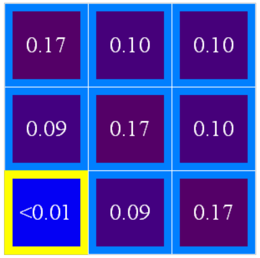
  - **Observed variables (evidence/fact)**: Agent knows **certain** things about the state of the world (e.g., sensor readings or symptoms or weather now)

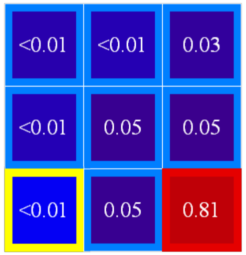
  - **Unobserved variables**: Agent needs to **reason** about **other** aspects (e.g. where an object is or what disease is present or weather tomorrow)

  - **Model**: Agent knows **something** about how the known variables **relate** to the unknown variables

- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

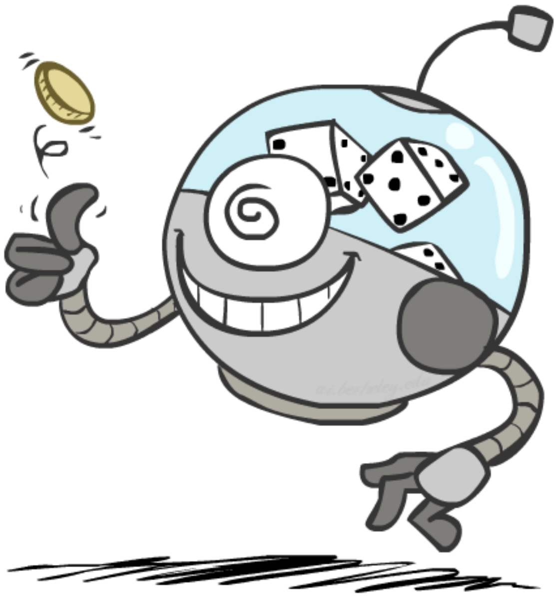






Random Variables

- A random variable is **some aspect of the world** about which we (may) have **uncertainty**
  - R = Is it raining?
  - T = Is it hot or cold?
  - D = How long will it take to drive to work?
  - L = Where is the ghost (Pacman project)?

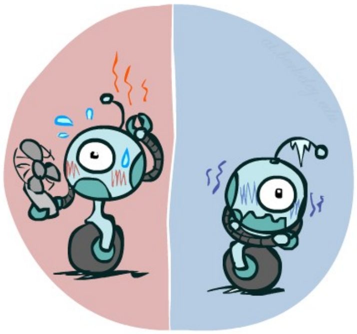


- We denote random variables with **capital letters**
- Random variables have **domains**
  - R in {true, false}   (often write as {+r, -r})
  - T in {hot, cold}
  - D in [0, ∞)
  - L in possible locations, maybe {(0,0), (0,1), …}

Probability Distributions

- Associate a probability with each value

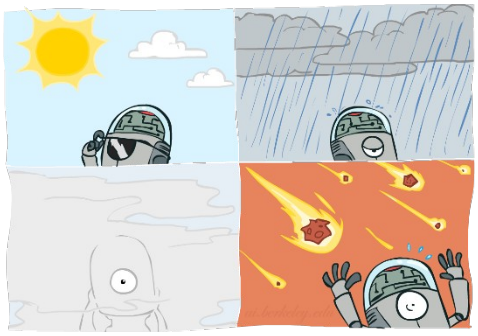
  - Temperature:



$P(T)$

T	P
hot	0.5
cold	0.5

  - Weather:



$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Probability Distributions

- Unobserved random variables have distributions

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Shorthand notation:

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$
$$\ldots$$

OK *if* all domain entries are unique

- A **distribution** is a **TABLE** of probabilities of values

- A **probability** (lower case value) is a **single number**

- Must have $P(W = rain) = 0.1$ and $\forall x \;\; P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Joint Distributions

- A *joint distribution* over **a set of random variables**: $X_1, X_2, \ldots X_n$
  specifies a real number for each *outcome*:

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

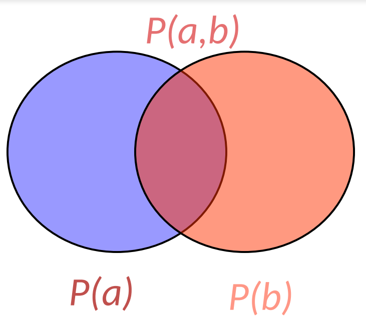
  - Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3



- Size of distribution if n variables with domain sizes d?
  - For all but the smallest distributions, impractical to write out

Events and Sample Space

- An *event* is a set E of outcomes

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of *any* event
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Example problem) Events

- P(+x, +y) ?
- P(+x) ?
- P(-y OR +x) ?

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Marginal Distributions

- Typically, the events we care about are *partial outcomes*, like P(T=hot)
- The set of **all possible outcomes** of an event E is its sample space
  - The sample space for it's hot = {it's hot and sun, it's hot and rain}
- Marginal distributions are **sub-tables** which **eliminate variables**
- Marginalization (summing out): Combine collapsed rows by adding

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_{w} P(t, w)$$

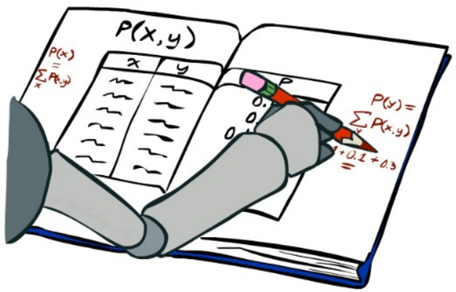
$$P(w) = \sum_{t} P(t, w)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4



In-class exercise: Marginal Distributions

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

$$P(x) = \sum_y P(x, y)$$

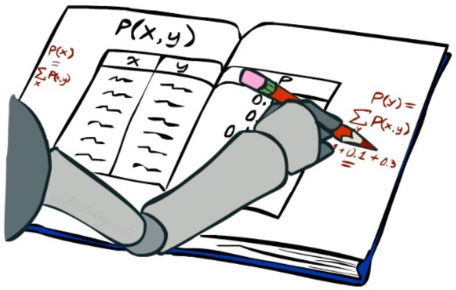
$$P(y) = \sum_x P(x, y)$$

$P(X)$

X	P
+x	
-x	

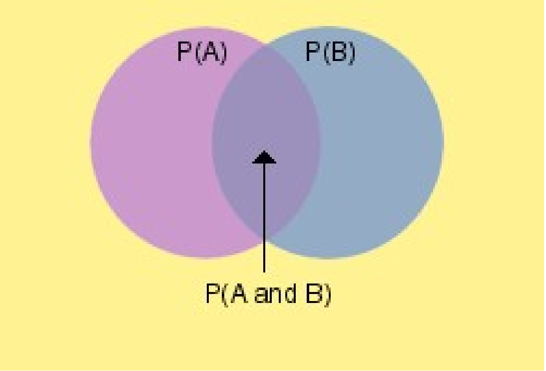
$P(Y)$

Y	P
+y	
-y	



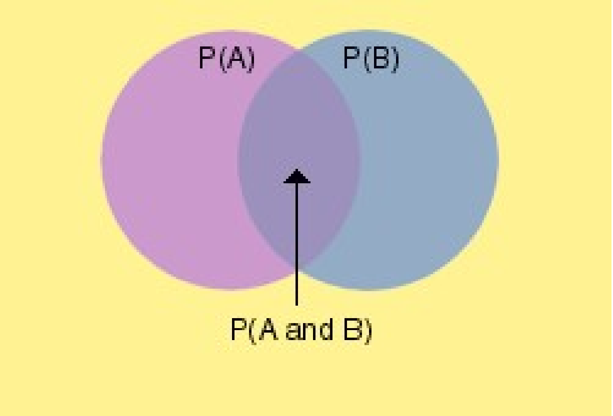
Conditional Probability

- **Probability** of an **event occurring** given that **another event** has **occurred** (by assumption, presumption, assertion or evidence)
- Notation:
  - –
  - Probability of event A, given B has already occurred
- Applications:
  - Medical diagnosis
    - “Probability of disease, given medical test is true”
    - Sensitivity of a test
    - Specificity of a test
  - Robot localization
    - Probability robot is at a specific location, given sensor values
  - Spam filtering
    - Probability email is spam, given email contains URLs and has spelling mistakes …



Conditional Probability Example

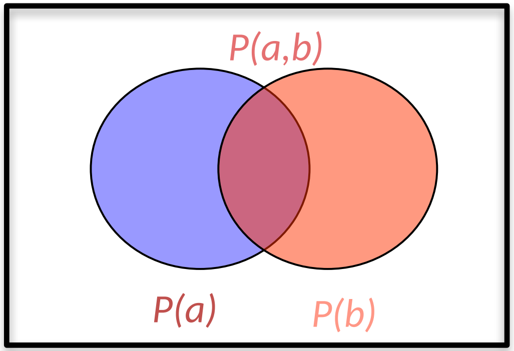
- Diagnosis using a clinical test
- Sample space = all patients tested
  - Event A: Subject has disease
  - Event B: Test is positive
- Interpret:
  - "probability of A and B"
    - Probability patient has disease and positive test (correct!)
    - Probability patient has disease BUT negative test (false negative)
    - Probability patient has no disease BUT positive test (false positive)
    - Probability patient has disease given a positive test
    - Probability patient has disease given a negative test



Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W = s, T = c) + P(W = r, T = c)$$
$$= 0.2 + 0.3 \quad = 0.5$$

In-class exercise –
Conditional Probabilities

- P(+x | +y) ?

- P(-x | +y) ?

- P(-y | +x) ?

$$P(X,Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Bayes' Rule

That's my rule!

- Two ways to factor a joint distribution over two variables: **p(x, y) = p(y, x)**

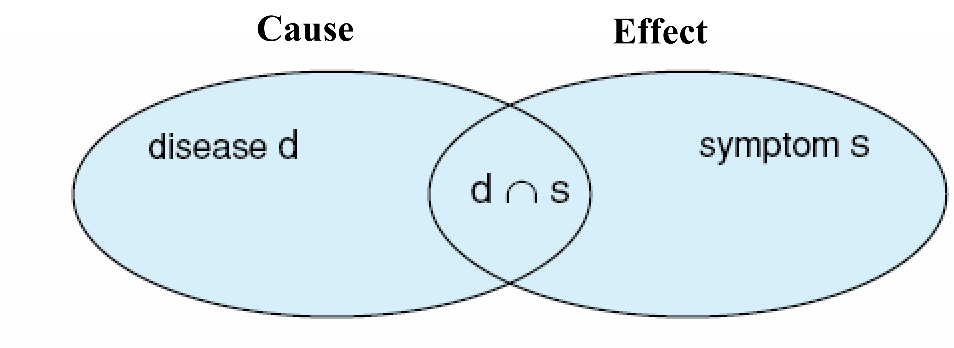
$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get: $P(x|y) = \dfrac{P(y|x)}{P(y)}P(x)$

- **Why is this at all helpful?**
  - Lets us build one conditional from its reverse
  - **Often one conditional is tricky but the other one is simple**
  - Foundation of many systems

- In the running for most important AI equation!

Interpretation of Bayes rule



Medical diagnosis system: study the relationship between disease and symptom

$$p(d|s) = \frac{p(s|d)p(d)}{p(s)}$$

Bayes' rule

p(d|s) means the probability of the disease **d** given symptom **s**. In many cases, it is difficult to know **p(d|s)**

If we know what the **disease** (or cause) is, telling those **symptoms** (or effect) of the disease, is much **easier** than figuring out a disease based on symptoms

s can be the result of a medical test

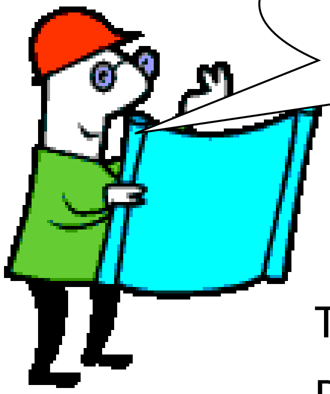
$p(s|+d)$ is test's sensitivity: TP/ (TP+FN)

$p(s|-d)$ is test's false positive rate: TP/ (TP+FN)

p(d) is occurrence of disease

Example Application of Bayes' Theorem

- Consider a manufacturing firm that receives shipment of parts from two suppliers.
- A_1 denotes the event that a part is received from supplier 1
- A_2 denotes the event that a part is received from supplier 2



We get 65 percent of our parts from supplier 1 and 35 percent from supplier 2.

Thus:

$P(A_1) = .65$ and $P(A_2) = .35$

Prior probabilities

Quality levels differ between suppliers

	Percentage Good Parts	Percentage Bad Parts
Supplier 1	98	2
Supplier 2	95	5

Let G denote that a part is good and B denote the event that a part is bad. Thus we have the following conditional probabilities:

$$P(G \mid A_1) = .98 \text{ and } P(B \mid A_1) = .02$$

$$P(G \mid A_2) = .95 \text{ and } P(B \mid A_2) = .05$$

Conditional probabilities

- A bad part broke one of our machines
- What is the probability the part came from supplier 1?

From the law of conditional probability:

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)}$$

Posterior probabilities?

$$P(A_1 \cap B) = P(A_1)P(B \mid A_1)$$

Tabular Approach to Bayes' Theorem—2-Supplier Problem

(1) Events A_i	(2) Prior Probabilities $P(A_i)$	(3) Conditional Probabilities $P(B \mid A_1)$	(4) Joint Probabilities $P(A_i \cap B)$	(5) Posterior Probabilities $P(A_i \mid B)$
A_1	.65	.02	.0130	.0130/.0305=.4262
A_2	.35	.05	.0175	.0175/.0305 =.5738
	1.00		P(B)=.0305	**1.0000**

In-class Exercise:
Numerical Example- “Cancer”

- Only 1 in 1000 people have **cancer**
- The True Positive Rate for a **cancer test** is 0.99
- The False Positive Rate for a **cancer test** is 0.02
- If one randomly tested individual is positive, what is the probability they have cancer?

Numerical Example

- Label events:
  - A = has disease A_o = no disease
  - B = Positive test result
- Examine probabilities
  - $p(A) = .001$
  - $p(A_o) = .999$
  - $p(B|A) = .99$
  - $p(B|A_o) = .02$

Tabular Approach to Bayes' Theorem—Cancer

(1) Events A_i	**(2)** Prior Probabilities $P(A_i)$	**(3)** Conditional Probabilities $P(B \mid A_i)$	**(4)** Joint Probabilities $P(A_i \cap B)$	**(5)** Posterior Probabilities $P(A_i \mid B)$
A				
A_0				

Naïve Bayes Classifier

- What if we have **multiple** tests to use to decide is a person has cancer?

- Simplified assumption: attributes are conditionally independent

Naïve Bayes Classifier

- Typically
  - a small number of outcomes or classes ()
  - conditional over several **feature variables** ()
- How to choose most likely class (one of the) given values of the feature variables?

  Naïve Bayes Classifier:

  Class =

  i.e., choose the class that has the highest posterior probability

- Note: is common to all probabilities
  - *No need to evaluate the denominator for comparisons*

Naïve Bayes Classifier: Example

Day	Outlook	Temperature	Humidity	Wind	*PlayTennis?*
1	Sunny	Hot	High	Light	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Light	Yes
4	Rain	Mild	High	Light	Yes
5	Rain	Cool	Normal	Light	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Light	No
9	Sunny	Cool	Normal	Light	Yes
10	Rain	Mild	Normal	Light	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Light	Yes
14	Rain	Mild	High	Strong	No

Naïve Bayes Classifier Example

- Problem: will I play tennis today given that it is:
  - Outlook=sunny
  - Temperature=cool
  - Humidity=high
  - Wind=strong
- P(PlayTennis | sunny, cool, high, strong)?

Naïve Bayes Classifier Example

- What are the events?
  - PlayTennis = yes
  - PlayTennis = no

- What are the features?
  - Outlook
  - Temperature
  - Humidity
  - Wind

Naïve Bayes Classifier

Use training data to calculate the conditional probabilities

P(PlayTennis=yes) = 9/14 = 0.64
P(PlayTennis=n) = 5/14 = 0.36
P(Outlook=sunny|PlayTennis=yes) = 2/9 = 0.22
P(Outlook=sunny|PlayTennis=no) = 3/5 = 0.60
P(Temperature=cool|PlayTennis=yes) = 3/9 = 0.33
P(Temperature=cool|PlayTennis=no) = 1/5 = .20
P(Humidity=high|PlayTennis=yes) = 3/9 = 0.33
P(Humidity=high|PlayTennis=no) = 4/5 = 0.80
P(Wind=strong|PlayTennis=yes) = 3/9 = 0.33
P(Wind=strong|PlayTennis=no) = 3/5 = 0.60

Naïve Bayes Classifier

P(PlayTennis=yes | sunny, cool, high, strong) $\propto$

P(PlayTennis=yes)xP(sunny|yes)xP(cool|yes)xP(high|yes)xP(strong|yes) = 0.0053

P(PlayTennis=no | sunny, cool, high, strong) $\propto$

P(PlayTennis=no)xP(sunny|no)xP(cool|no)xP(high|no)xP(strong|no) = 0.0206

So the class for this instance is PlayTennis=NO:

We are more likely to ***not*** play tennis as compared to playing tennis

We can normalize the probability by:

P(PlayTennis=no | sunny, cool, high, strong) = = 0.795

In-class Exercise:
Naïve Bayes Classifier

Will I play tennis today given that it is:

- Outlook=sunny
- Temperature=hot
- Humidity=normal
- Wind=strong

Naïve Bayes Classifier Example

- Problem: will I play tennis today given that it is:
  - Outlook=overcast
  - Temperature=cool
  - Humidity=high
  - Wind=strong

Naïve Bayes Classifier

P(PlayTennis=yes) = 9/14 = 0.64

P(PlayTennis=no) = 5/14 = 0.36

P(Outlook=overcast|PlayTennis=yes) = 4/9 = 0.44

P(Outlook=overcast|PlayTennis=no) = 0/5 = 0

P(Temperature=cool|PlayTennis=yes) = 3/9 = 0.33

P(Temperature=cool|PlayTennis=no) = 1/5 = .20

P(Humidity=high|PlayTennis=yes) = 3/9 = 0.33

P(Humidity=high|PlayTennis=no) = 4/5 = 0.80

P(Wind=strong|PlayTennis=yes) = 3/9 = 0.33

P(Wind=strong|PlayTennis=no) = 3/5 = 0.60

Naïve Bayes Classifier

Estimating Probabilities:

- In the previous example, P(overcast|no) = 0 which causes the formula-
- P(no)xP(overcast|no)xP(cool|no)xP(high|no)xP(strong|no) = 0.0
- This causes problems in comparing because the other probabilities are not considered.
- There are methods to avoid this -> Additive (Laplace) smoothing

$$P(feature|class) = \frac{(\text{number of times feature appears with class})+1}{(\text{number of instances of class})+k}$$

Naïve Bayes Classifier in practice

- Classification is efficient
- The conditional probability values of all the attributes with respect to the class are pre-computed and stored on disk.
- No need to re-compute the conditional probabilities for a new problem instance

Application: Email spam filtering

- Spam: unsolicited bulk email
- Large fraction of all email sent
  - Estimates 64-77%
- Spam filtering
  - Feature-based
  - Applied at receiver
  - Email client, mail server, or ISP

Send | Options... | Help

To... hiring@123publishing.com

Cc...

Bcc...

Subject: Editorial Assistant Position - Susan Sharp

Attachments:

Normal | Arial | 10

Dear Hiring Manager,

I would like to express my interest in a position as editorial assistant for your publishing company. As a recent graduate with writing, editing, and administrative experience, I believe I am a strong candidate for a position at the 123 Publishing Company.

You specify that you are looking for someone with strong writing skills. As an English major, a writing tutor, and an editorial intern for both a government magazine and a college marketing office, I have become a skilled writer with a variety of experience.

Although I am a recent college graduate, my maturity, practical experience, and eagerness to enter the publishing business will make me an excellent editorial assistant. I would love to begin my career with your company, and am confident that I would be a beneficial addition to the 123 Publishing Company.

I have attached my resume. Thank you so much for your time and consideration.

Sincerely,

Susan Sharp

———

Susan Sharp
123 Main Street
XYZ Town, NY 11111
Email: susan.sharp@mail.com
Cell: 555-555-5555

googleteam GOOGLE LOTTERY WINNER! CONTACT

From: googleteam **To:**
Subject: GOOGLE LOTTERY WINNER! CONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

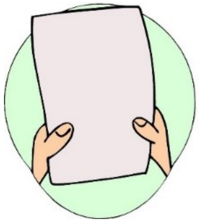
GOOGLE LOTTERY INTERNATIONAL
INTERNATIONAL PROMOTION / PRIZE AWARD .
(WE ENCOURAGE GLOBALIZATION)
FROM: THE LOTTERY COORDINATOR,
GOOGLE B.V. 44 9459 PE.
RESULTS FOR CATEGORY "A" DRAWS
Congratulations to you as we bring to your notice, the results of the First Ca
inform you that your email address have emerged a winner of One Million (1,(
money of Two Million (2,000,000.00) Euro shared among the 2 winners in this
email addresses of individuals and companies from Africa, America, Asia, Au
CONGRATULATIONS!
Your fund is now deposited with the paying Bank. In your best interest to avo
award strictly from public notice until the process of transferring your claims
NOTE: to file for your claim, please contact the claim department below on e
****

Classification for Spam Filtering

- Extract features from header, content
- Train Naïve Bayes classifier
- Classify message:
  - Spam or not

- Define classes:

vs.



Features

- Convert message to ***feature vector***
- Where do we get features?
  - From header
    - sender, recipient, routing,…
    - Possibly break up domain names
  - From content
    - Words, phrases, character strings
    - Become binary or numeric features
  - Presence/absence of URLs, HTML tags, images,…

Conditional probabilities

- From a collection of known spam/genuine emails, calculate:
  - p(senderName=A|spam)
  - p(senderDomainName=aol.com|spam)
  - p(usernameContainsNumbers=true|spam)
  - p(contentContainsExclamation=true |spam)
  -
  -
  -
- Also, need to calculate prior probabilities
  - p(spam) = 0.7

Classifying as spam

- Given a new email message, calculate:
  1. senderDomainName=aol.com?
  2. usernameContainsNumbers=true?
  3. contentContainsExclamation=true?
  4. …
  5.
  6.
- The feature vector
- Calculate
  - p(spam|feature vector)=p(spam)p(F1|spam)p(F2|spam)…p(Fn|spam)
  - If p(spam|feature vector) > p(not_spam|feature vector) then Spam

Numerical example for spam filtering

A Naïve Bayes classifier is used for detecting if email is spam or not using the following three features:

Feature 1: Sender domain is aol.com

Values = yes, no

p(F1 = yes | email=spam) = 0.4

p(F1 = yes | email=not spam) = 0.2

Feature 2: Sender username contains numbers

Values = yes, no

p(F2 = yes | email=spam) = 0.8

p(F2 = yes | email=not spam) = 0.4

Feature 3: Subject contains exclamation points (contains “!”)

Values = yes, no

p(F3 = yes | email=spam) = 0.8

p(F3 = yes | email= not spam) = 0.1

The prior probability that an email is spam is 70% (p(email=spam) = 0.70)

Numerical example for spam filtering

Consider the following two email segments:

- Give the feature vectors for the above two emails.

Email 1:
```
To: user@fullerton.edu
From: george123@aol.com
Subject: Congratulations!!! you
have won the state lottery!
Date: 1 Oct 2015, 23:27:09
+0000 (UTC)
```

Email 2:
```
To: user@fullerton.edu
From: alan.smith@aol.com
Subject: Meeting at 2pm
Date: 12 Nov 2015,
05:23:49 +0000 (UTC)
```

Numerical example for spam filtering

Consider the following two email segments:

- Give the feature vectors for the above two emails.

Email 1:
```
To: user@fullerton.edu
From: george123@aol.com
Subject: Congratulations!!! you
have won the state lottery!
Date: 1 Oct 2015, 23:27:09
+0000 (UTC)
```

Email 2:
```
To: user@fullerton.edu
From: alan.smith@aol.com
Subject: Meeting at 2pm
Date: 12 Nov 2015,
05:23:49 +0000 (UTC)
```

	Feature 1	Feature 2	Feature 3
Email 1	yes	yes	yes
Email 2	yes	no	no

Email 1

- Spam
  - P(spam)*p(+f1|spam)*p(+f2|spam)*p(+f3|spam)
  - =0.7 * 0.4 * 0.8 * 0.8 = 0.1792
- Not Spam
  - P(-spam)*p(+f1|-spam)*p(+f2|-spam)*p(+f3|-spam)
  - =(1-0.7) * 0.2 * 0.4 * 0.1 = 0.0024
- Spam
- P(Spam|+f1,+f2,+f3) = 0.1792/(0.1792+0.0024) = 0.986 = 98.6%

Naïve Bayes advantages

- Independence allows parameters to be estimated on different data sets, e.g.
  - Estimate content features from messages with headers omitted
  - Estimate header features from messages with content missing

Randomly generated name and email

From: Sam Elegy <aj6xfdou7@yahoo.com>
To: ddlewis4@att.net
Subject: you can buy V!@gra

Typographic variations

No doctor visit needed

Spam like content in images

I don't like emails.
than named did the and people other FINDS for itself of to such the U.S. liberty gives enforced Bureau Civil Constitution, published he judge House NEW allowing public the Civil

Irrelevant content

Defeating Feature Extraction

- Misspellings, character set choice, HTML games: mislead extraction of words
- Put content in images
- Forge headers (to avoid identification, but also interferes with classification)
- Non-spam content to mimic distribution in nonspam
- “Hashbusters”
  - Try to defeat Bayesian spam filtering by making each individual spam look as different as possible

References

- George F. Luger, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 6th edition, Addison Wesley, 2009. **Chapters 4.1.1, 12.1**.
- Russel and Norvig, Artificial Intelligence: A Modern Approach, 3rd edition, Prentice Hall, 2010. **Chapter 4-4.1**