Data Mining - evaluation
Suppose we want to know if a person has a specific cancer or not.

Suppose we have a test that will return + or -

We have prior knowledge that over the entire population only .008 (.8%) have this disease.

The test returns a correct positive 98% of the time a person has cancer.

It returns a correct negative 97% of the time a person doesn’t
Suppose we want to know if a person has a specific cancer or not.

Suppose we have a test that will return + or -

We have prior knowledge that over the entire population only .008 (.8%) have this disease.

The test returns a correct positive 98% of the time a person has cancer

It returns a correct negative 97% of the time a person doesn’t

What is the probability the next slide will be about Naive Bayes?
The test returns a correct positive 98% of the time a person has cancer.

It returns a correct negative 97% of the time a person doesn’t.

TP
TN
FP
FN
The test returns a correct positive 98% of the time a person has cancer.

It returns a correct negative 97% of the time a person doesn’t.

TP= ??
TN= ??
FP= ??
FN= ??
The test returns a correct positive 98% of the time a person has cancer.

It returns a correct negative 97% of the time a person doesn’t.

TP = .98
TN = .97
FP = .03
FN = .02
The test returns a correct positive 98% of the time a person has cancer

It returns a correct negative 97% of the time a person doesn’t

Sample size of 200 - 100 w/ cancer 100 w/o

TP= .98
TN= .97
FP= .03
FN=.02
The test returns a correct positive 98% of the time a person has cancer.

It returns a correct negative 97% of the time a person doesn’t.

Sample size of 200 - 100 w/ cancer 100 w/o

<table>
<thead>
<tr>
<th>TP</th>
<th>Test says positive</th>
<th>Test says negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN</td>
<td>.97</td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td>FN</td>
<td>.02</td>
<td></td>
</tr>
</tbody>
</table>

Real world positive
Real world negative
The test returns a correct positive 98% of the time a person has cancer.

It returns a correct negative 97% of the time a person doesn’t.

Sample size of 200 - 100 w/ cancer 100 w/o

<table>
<thead>
<tr>
<th>Test says positive</th>
<th>Test says negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real world positive</td>
<td>TP = 98</td>
</tr>
<tr>
<td>Real world negative</td>
<td>FP=3</td>
</tr>
<tr>
<td></td>
<td>FN=2</td>
</tr>
<tr>
<td></td>
<td>TN=97</td>
</tr>
</tbody>
</table>

TP= .98
TN= .97
FP= .03
FN=.02
The test returns a correct positive 98% of the time a person has cancer.

It returns a correct negative 97% of the time a person doesn’t.

Sample size of 200 - 100 w/ cancer 100 w/o

What is the accuracy of the test?

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%
\]

\[
\frac{.98 + .97}{.98 + .97 + .03 + .02} = 97.5\%
\]
• The test returns a correct positive 98% of the time a person has cancer

• It returns a correct negative 97% of the time a person doesn’t

Sample size of 200 - 100 w/ cancer 100 w/o

What is the accuracy of the test?

TP= .98
TN= .97
FP= .03
FN=.02

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%
\]

97.5%

Are FP and FN equal in importance?
8 people out of 1000 have Pyruvate kinase deficiency

Test 1
TP=.999
TN=.990
FP=.01
FN=.001
accuracy=99.5

Test 2
TP=.990
TN=.990
FP=.001
FN=.01
accuracy=99.5

Test 1 and 2 mutually exclusive
Without treatment, a person will die in 24 hours. Treatment is taking Azizafened, a drug with minor side effects. Which test to administer?

Test 1
TP=.999
TN=.990
FP=.01
FN=.001
accuracy=99.5

Test 2
TP=.990
TN=.990
FP=.001
FN=.01
accuracy=99.5

Test 1 and 2 mutually exclusive
Test 1
TP=.999
TN=.990
FP=.01
FN=.001
accuracy=99.5

Test 2
TP=.990
TN=.990
FP=.001
FN=.01
accuracy=99.5

Test says
positive

Test says
negative

Real world
positive
800

Real world
negative
99,200

100,000 people

TP =
FN =
FP =
TN =
Test 1
TP=.999
TN=.990
FP=.01
FN=.001
accuracy=99.5

Test 2
TP=.990
TN=.990
FP=.001
FN=.01
accuracy=99.5

Test says positive
Test says negative

Real world positive
TP1=799
TP2=792
FN1=1
FN2=8

Real world negative
FP1=992
FP2=92
TN1=98,208
TN2=99,108

100,000 people

800
99,200
<table>
<thead>
<tr>
<th>100,000 people</th>
<th>Test says positive</th>
<th>Test says negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real world positive</td>
<td>TP1 = 799</td>
<td>FN1 = 1</td>
</tr>
<tr>
<td>800</td>
<td>TP2 = 792</td>
<td>FN2 = 8</td>
</tr>
<tr>
<td>Real world negative</td>
<td>FP1 = 992</td>
<td>TN1 = 98,208</td>
</tr>
<tr>
<td>99,200</td>
<td>FP2 = 92</td>
<td>TN2 = 99,108</td>
</tr>
</tbody>
</table>

Which cell is the one where people die?
<table>
<thead>
<tr>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP=.999</td>
<td>TP=.990</td>
</tr>
<tr>
<td>TN=.990</td>
<td>TN=.990</td>
</tr>
<tr>
<td>FP=.01</td>
<td>FP=.001</td>
</tr>
<tr>
<td>FN=.001</td>
<td>FN=.01</td>
</tr>
<tr>
<td>accuracy=99.5</td>
<td>accuracy=99.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>100,000 people</th>
<th>Test says positive</th>
<th>Test says negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>800 Real world positive</td>
<td>TP1 =799</td>
<td>FN1=1</td>
</tr>
<tr>
<td></td>
<td>TP2=792</td>
<td>FN2=8</td>
</tr>
<tr>
<td>99,200 Real world negative</td>
<td>FP1=992</td>
<td>TN1=98,208</td>
</tr>
<tr>
<td></td>
<td>FP2=92</td>
<td>TN2=99,108</td>
</tr>
</tbody>
</table>

Change story

8 out of 1000 are compatible w/ any one person wrt kidney transplants
If test negative, person removed from candidate pool. If test is positive, person subject to further painful testing which involves long needles.

100,000 people

800
Real world positive
TP1 = 799
TP2 = 792

99,200
Real world negative
FP1 = 992
FP2 = 92
Summary

just an accuracy measure may not be sufficient.

Need to look at FP, FN, and see how that matches problem.
Which to minimize: FP or FN?

- I own a startup and am looking to hire 5 new developers.
- There are 300k layoffs this year so the pool of candidates is fantastic.
- What I don’t want is to hire a loser
- I have various hiring methods I can use, each w/ a known FP and FN. Which should I select?
Which to minimize

- I own a roofing company
- I develop a data mining tool to select people who are likely to buy our service.
- With the people identified by the tool, my sales force will call them to do direct marketing.
Which to minimize

- a smoke detector
- pandora recommender
- terrorist detector
Related Concepts- Recall/Precision

- **Precision**: Percentage of retrieved documents that are relevant: $\frac{TP}{TP+?}$

- **Recall**: Percentage of relevant documents that are returned: $TP(TP + ?)$
Related Concepts- Recall/Precision

- **Precision**: Percentage of retrieved documents that are relevant: \( \frac{TP}{TP+FP} \)

- **Recall**: Percentage of relevant documents that are returned: \( TP / (TP + FN) \)
What about browsing and search?
What about browsing and search?

- Search results don’t have to be very good
- Recall? Not important. as long as you get at least some good hits.
- Precision? Not important. As long as at least some of the hits on the first page you return are good.
Docker - ArchWiki - Arch Linux
https://wiki.archlinux.org/index.php/Docker  Arch Linux
Jump to Build Image - Instead, check docker base/archlinux registry and click the mkimage-arch.sh link to download mkimage-arch.sh and ...

Official 13.1 Docker Containers Released - openSUSE News
news.opensuse.org › 2014 › 08 › 07  openSUSE
Aug 7, 2014 - Creating a docker application based on the official containers is easy too. ... It creates one of the world's best Linux distributions, working ...

Oracle Linux images for Docker released (Oracle's Linux Blog)
https://blogs.oracle.com/linux/.../oracle_linux_images...  Oracle Corporation
Nov 13, 2014 - The Docker Engine and Official Oracle Linux images are fully supported as part of Oracle Linux Basic and Premier support subscriptions.

Smaller Java images with Alpine Linux - Atlassian Developers
https://developer.atlassian.com/.../minimal-java-docker-containe...  Atlassian
Aug 10, 2015 - docker build -t yourname/minimal-java . ... docker run -t durdn/minimal-java ... Install curl , tar , and ca-certificates on the base alpine image.

Building Docker Images — Project Atomic
www.projectatomic.io/docs/docker-building-images/  Project Atomic
Resource management and process isolation come from Linux Containers (LXC). ... The Dockerfile describes a base image for the build using the FROM ...

How To Install and Use Docker: Getting Started | DigitalOcean
https://www.digitalocean.com/.../how-to-install-and-use-docker-getting-st...  DigitalOcean
Dec 11, 2013 - docker containers: directories containing everything-your-application; docker images: snapshots of containers or base OS (e.g. Ubuntu) images ...
Size of Docker images: Which linux is smaller? | Java Bien!
blog.javabien.net/2014/06/.../size-of-docker-images-which-linux-is-small... ▼
Jun 18, 2014 - That means that if you use multiple images based on the same base ...
and an ubuntu:14.04 containers to tar files using docker export. docker ...

Docker Quick Start | SUSE Linux Enterprise Server 12
https://www.suse.com/.../sles-12/.../dockerquick/dockerquick.html ▼ SUSE ▼
Oct 23, 2015 - 1 Terminology; 2 Overview; 3 Setting Up a Docker Host; 4 Basic Docker
Operations; 5 Building Docker Images; 6 SUSE Linux Enterprise ...

AWS Developer Forums: Amazon Linux as docker container ...
May 4, 2014 - 12 posts
Pretty lame that AWS linux is touted as having docker support etc, etc, but yet there is
no official AWS linux base image? Whats up with this?

Creating an Oracle Enterprise Linux 6 base image for Docker
blog.grid-it.nl/.../creating-an-oracle-enterprise-linux-6-base-image-for-d... ▼
May 10, 2014 - In this post we will create an Oracle Enterprise Linux 6 base image,
which we can use to create other images, for example to run Oracle XE in.
a student is doing a paper on the intestinal absorption (bioavailability) of calcium. what is the precision and recall of these search results?

- Calcium does not inhibit iron absorption or alter iron status in infant piglets adapted to a high calcium diet.
- Calcium absorption from small soft-boned fish.
- Bioavailability and pharmacokinetic characteristics of two monofluorophosphate preparations with calcium supplement.
- Effects of milk and milk components on calcium, magnesium, and trace element absorption during infancy.
- Bioavailability of calcium and magnesium.
- The inhibitory effect of dietary calcium on iron bioavailability: a cause for concern?
- Iron absorption from the whole diet: comparison of the effect of two different distributions of daily calcium intake.
- Strontium as a marker for intestinal calcium absorption: the stimulatory effect of calcitriol.
- Calcium bioavailability and parathyroid hormone acute changes after oral intake of dairy and nondairy products in healthy volunteers.
- Intestinal absorption of calcium from foodstuffs as compared to a pharmaceutical preparation.
- The bioavailability of dietary calcium.
- Searching for the determinants of intestinal calcium absorption
Precision and Recall - an inverse relationship

How can I maximize recall?
How can I maximize precision?
Other problems

Items exist that are marginally relevant.
There are degrees of relevance.
amazon

soy extract

Date
calculate the FP & FN for a few classification methods on the Pima data and/or the cancer data.

Recall that the diabetes data is if the person will get diabetes in one year. The cancer data set is whether the tumor is malignant or benign.

For these 2 data sets which (if any) measure is more important?