CS440/ECE448 Lecture 3: Decision Theory

Mark Hasegawa-Johnson

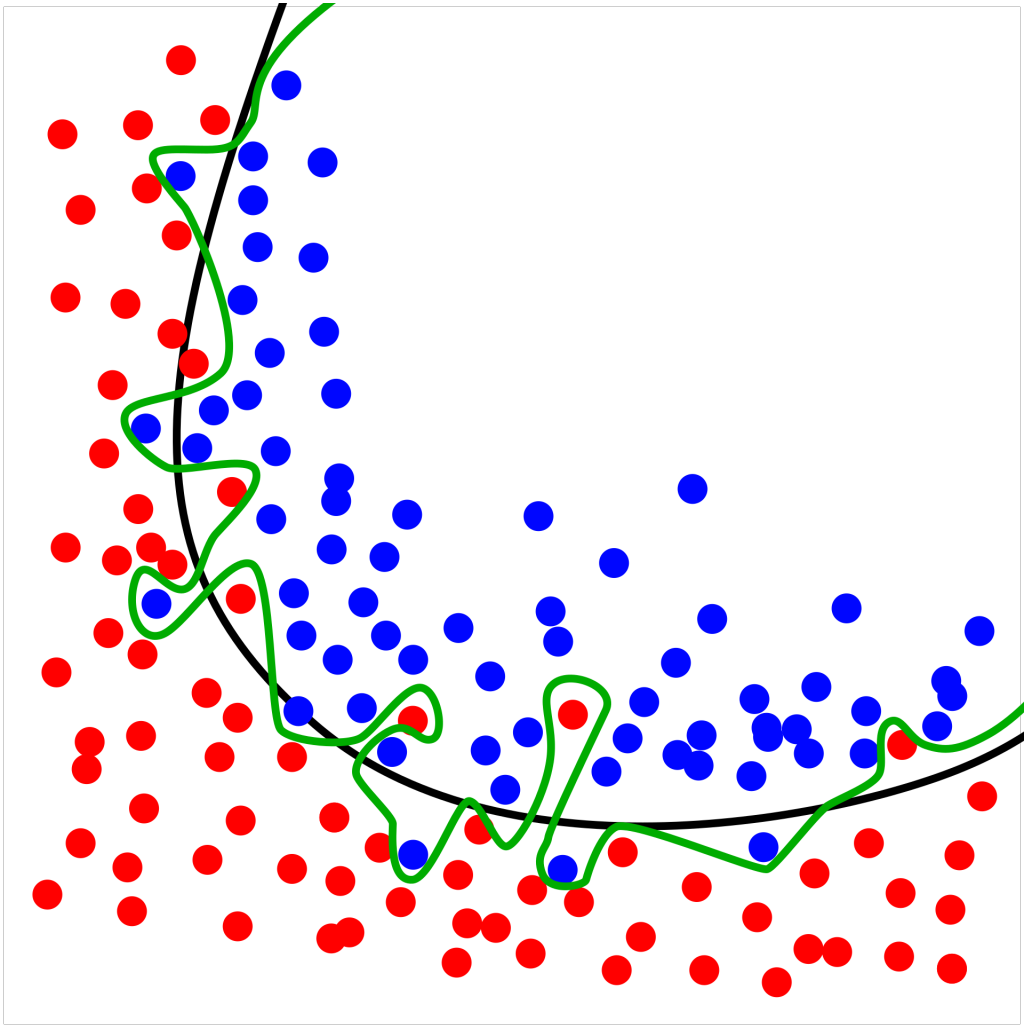


Diagram showing overfitting of a classifier. CC-BY 4.0, Ignacio Icke, 2008

Outline

- Decision Theory
- Minimum Probability of Error
- Bayes' Rule
- Accuracy, Error Rate, and the Bayes Error Rate
- Confusion Matrix, Precision & Recall, Sensitivity & Specificity

Today's example: Spam filter

- Junk e-mail is called "spam"
- Useful e-mail is called "ham."
- In *The Spam Song*, by the comedian troupe Monty Python, a restaurant claims to offer anything you want, but in fact, serves nothing but spam
- The internet claims to offer anything you want, but in fact, offers nothing but spam

https://en.wikipedia.org/wiki/File:SPAM_SONG_7%22.jpg

Decision Theory

- Suppose we have an experiment with two random variables, X and Y.
  - X is something we can observe, like the words in an email.
  - Y is something we can't observe, but we want to know. For example, Y=1 means the email is spam, Y=0 means it's ham (desirable mail).
- Can we train an AI to read the email, and determine whether it's spam or not?

Decision Theory

- Y = the correct label
  - Y = the correct label as a random variable ("in general")
  - y = the label observed in a particular experiment ("in particular")
- $f(X)$ = the decision that we make, after observing the datum, X
  - $f(X)$ = the function applied to random variable X ("in general")
  - $f(x)$ = the function applied to a particular value of x ("in particular")

Deciding how to Decide: Loss and Risk

- Suppose that deciding $f(x)$, when the correct label is $Y = y$, costs us a certain amount of money (or prestige, or safety, or points, or whatever) – call that the **loss**, $l(f(x), y)$
- In general, we would like to lose as few points as possible (negative losses are good…)
- Define the **risk**, $R(f)$, to be the expected loss incurred by using the decision rule $f(X)$:

$$R(f) = E[l(f(X), Y)] = \sum_y \sum_x l(f(x), y) P(X = x, Y = y)$$

Minimum-Risk Decisions

- If we want to the smallest average loss (the smallest risk), then our decision rule should be

$$f = \operatorname{argmin} R(f)$$

- In other words, for each possible x, we find the value of $f(x)$ that minimizes our expected loss given that x, and that is the $f(x)$ that our algorithm should produce.

Outline

- Decision Theory
- Minimum Probability of Error
- Bayes' Rule
- Accuracy, Error Rate, and the Bayes Error Rate
- Confusion Matrix, Precision & Recall, Sensitivity & Specificity

Zero-One Loss

Suppose that $f(x)$ is an estimate of the correct label, and

- We lose one point if $f(x) \neq y$
- We lose zero points if $f(x) = y$

$$l(f(x), y) = \begin{cases} 1 & f(x) \neq y \\ 0 & f(x) = y \end{cases}$$

Then the risk is just the probability of error:

$$R(f) = E[l(f(X), Y)] = \Pr(f(X) \neq Y)$$

Minimum Probability of Error

We can minimize the probability of error by designing f(x) so that $f(x) = 1$ when $Y = 1$ is more probable, and $f(x) = 0$ when $Y = 0$ is more probable.

$$f(x) = \begin{cases} 1 & P(Y = 1|X = x) > P(Y = 0|X = x) \\ 0 & P(Y = 1|X = x) < P(Y = 0|X = x) \end{cases}$$

What other types of risk matter?

- Risk = Expected loss
- For zero-one loss, Risk=probability of error
- ”Probability of error” is one type of risk you might try to minimize, but it’s not the only one. What other types of risk might be worth minimizing, for what types of problems?

Outline

- Decision Theory
- Minimum Probability of Error
- Bayes' Rule
- Accuracy, Error Rate, and the Bayes Error Rate
- Confusion Matrix, Precision & Recall, Sensitivity & Specificity

Example: spam detection

How can we estimate $P(Y = y|X = x)$?

- The prior probability of spam might be obvious. If 80% of all email on the internet is spam, that means that

$$P(Y = 1) = 0.8, P(Y = 0) = 0.2$$

- The probability of X given Y is also easy. Suppose we have a database full of sample emails, some known to be spam, some known to be ham. We count how often any word occurs in spam vs. ham emails, and estimate:

$$P(X = x|Y = 1) = \text{fraction of the words in spam emails that are the word } x$$
$$P(X = x|Y = 0) = \text{fraction of the words in ham emails that are the word } x$$

- Now we have $P(X = x|Y = y)$ and $P(Y = y)$. How do we get $P(Y = y|X = x)$?

Bayes' Rule

Rev. Thomas Bayes
(1702-1761)

The reverend Thomas Bayes solved this problem for us in 1763. His proof is really just the definition of conditional probability, applied twice in a row:

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

$$= \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

The four Bayesian probabilities

$$P(Y = y|X = x) = \frac{P(Y = y)P(X = x|Y = y)}{P(X = x)}$$

This equation shows the relationship among four probabilities. This equation has become so world-famous, since 1763, that these four probabilities have standard universally recognized names that you need to know:

- $P(Y = y|X = x)$ is the ***a posteriori*** (after-the-fact) probability, or **posterior**
- $P(Y = y)$ is the ***a priori*** (before-the-fact) probability, or **prior**
- $P(X = x|Y = y)$ is the **likelihood**
- $P(X = x)$ is the **evidence**

Bayes' rule applied

- If we know the prior and the likelihoods:

$$P(Y = y) = \text{fraction of emails that are of type } y$$

$$P(X = x|Y = y) = \text{fraction of the words in } y \text{ emails that are the word } x$$

- …then the posterior is:

$$P(Y = y|X = x) = \frac{P(Y = y)P(X = x|Y = y)}{P(X = x)}$$

- … and the minimum-probability-of-error classifier is:

$$f(x) = \underset{y}{\operatorname{argmax}}\, P(Y = y|X = x) = \underset{y}{\operatorname{argmax}}\, P(Y = y)P(X = x|Y = y)$$

Minimum probability of error, redux

The minimum-probability-of-error classifier is:

$$f(x) = \underset{y}{\operatorname{argmax}}\, P(Y = y | X = x) = \underset{y}{\operatorname{argmax}}\, P(Y = y, X = x)$$

This is sometimes called the “Bayesian classifier.”

Outline

- Decision Theory
- Minimum Probability of Error
- Bayes' Rule
- Accuracy, Error Rate, and the Bayes Error Rate
- Confusion Matrix, Precision & Recall, Sensitivity & Specificity

Accuracy

When we train a classifier, the metric that we usually report is "accuracy."

$$\text{Accuracy} = \frac{\#\text{ tokens correctly classified}}{\#\text{ tokens total}}$$

Error Rate

Equivalently, we could report error rate, which is just 1-accuracy:

$$\text{Error Rate} = \frac{\#\text{ tokens incorrectly classified}}{\#\text{ tokens total}}$$

Bayes Error Rate

The “Bayes Error Rate” is the smallest possible error rate of any classifier with labels y and features x:

$$\text{Error Rate} = \sum_{x} P(X = x) \min_{y} P(Y \neq y | X = x)$$

It’s called the “Bayes error rate” because it’s the error rate of the Bayesian classifier.

Outline

- Decision Theory
- Minimum Probability of Error
- Bayes' Rule
- Accuracy, Error Rate, and the Bayes Error Rate
- Confusion Matrix, Precision & Recall, Sensitivity & Specificity

The problem with accuracy

- In most real-world problems, there is one class label that is much more frequent than all others.
  - Words: most words are nouns
  - Animals: most animals are insects
  - Disease: most people are healthy
- It is therefore easy to get a very high accuracy. All you need to do is write a program that completely ignores its input, and always guesses the majority class. The accuracy of this classifier is called the "chance accuracy."
- It is sometimes very hard to beat the chance accuracy. If chance=90%, and your classifier gets 89% accuracy, is that good, or bad?

The solution: Confusion Matrix

<u>Confusion Matrix</u> =

- $(m, n)^{\text{th}}$ element is
- the number of tokens of the m^{th} class
- that were labeled, by the classifier, as belonging to the n^{th} class.

```
title: Consonant Confusions in CV utterances, for V=/a/, for S/N = +12db and
Phones involved: 16, namely p t k f T (th) s S (sh) b d g v D (dh) z Z (zh) m

   p   t   k   f   T   s   S   b   d   g   v   D   z   Z   m   n      Total
p  228 7   7   1   0   0   1   0   0   0   0   0   0   0   0   0    p 244
t  0   236 8   0   0   0   0   0   0   0   0   0   0   0   0   0    t 244
k  26  5   213 0   0   0   0   0   0   0   0   0   0   0   0   0    k 244
f  6   1   1   194 35  0   0   3   0   0   1   3   0   0   0   0    f 244
T  0   2   2   96  146 2   0   2   1   0   1   8   0   0   0   0    T 260
s  0   2   0   1   31  204 1   1   9   4   0   7   0   0   0   0    s 260
S  0   0   0   0   0   1   243 0   0   0   0   0   0   0   0   0    S 244
b  0   0   0   13  12  0   0   207 2   3   19  8   0   0   0   0    b 264
d  0   0   0   0   0   0   0   0   240 9   0   0   0   3   0   0    d 252
g  0   0   0   0   0   0   0   1   41  199 0   0   2   1   0   0    g 244
v  0   0   0   3   3   0   0   20  0   2   182 47  2   0   0   1    v 260
D  0   0   0   0   7   0   0   10  3   22  49  170 19  0   0   0    D 280
z  0   0   0   0   1   0   0   3   8   24  2   22  145 3   0   0    z 208
Z  0   0   0   0   0   0   1   0   2   0   0   0   13  264 0   0    Z 280
m  0   0   0   0   0   0   0   0   0   0   0   0   0   0   213 11   m 224
n  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   248  n 248
```

Plaintext versions of the Miller & Nicely matrices, posted by Dinoj Surendran, http://people.cs.uchicago.edu/~dinoj/research/nicely.html

Confusion matrix for a binary classifier

Suppose that the correct label is either 0 or 1. Then the confusion matrix is just 2x2.

For example, in this box, you would write the # tokens of class 1 that were misclassified as class 0

False Positives & False Negatives

- TP (True Positives) = tokens that were correctly labeled as “1”
- FN (False Negatives) = tokens that should have been “1”, but were mislabeled as “0”
- FP (False Positives) = tokens that should have been “0”, but were mislabeled as “1”
- TN (True Negative) = tokens that were correctly labeled as “0”

Classified As:

Correct Label:	0	1
0	TN	FP
1	FN	TP

Summaries of a Binary Confusion Matrix

The binary confusion matrix is standard in many fields, but different fields summarize its content in different ways.

- In medicine, it is summarized using Sensitivity and Specificity.
- In information retrieval, it is summarized using Recall and Precision.

Classified As:

Correct Label:	0	1
0	TN	FP
1	FN	TP

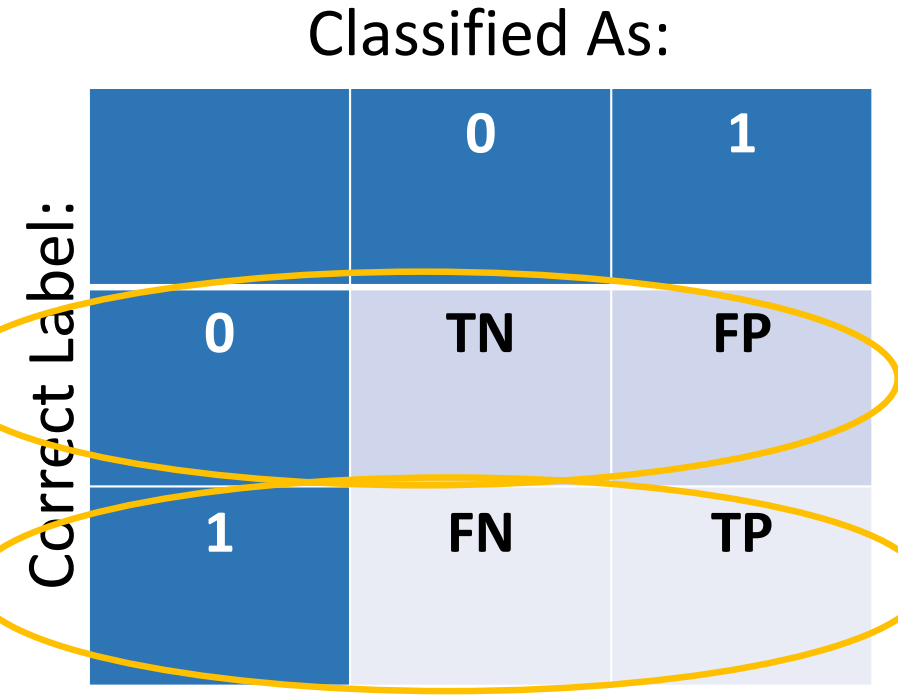
Medicine: Specificity and Sensitivity

Specificity = True Negative Rate (TNR):

$$TNR = P(f(X) = 0|Y = 0) = \frac{TN}{TN + FP}$$

Sensitivity = True Positive Rate (TPR):

$$TPR = P(f(X) = 1|Y = 1) = \frac{TP}{TP + FN}$$



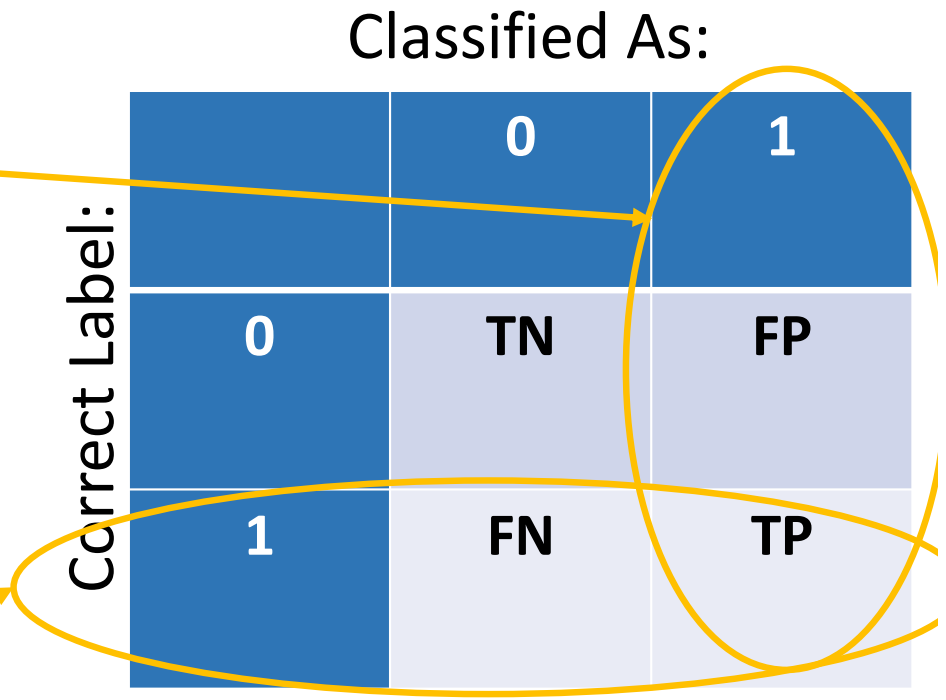
Information Retrieval: Recall & Precision

Precision:

$$P = P(Y = 1 | f(X) = 1) = \frac{TP}{TP + FP}$$

Recall = Sensitivity = TPR:

$$R = P(f(X) = 1 | Y = 1) = \frac{TP}{TP + FN}$$



The Misdiagnosis Problem: Example

1% of women at age forty who participate in routine screening have breast cancer. The test has sensitivity of 80%, and selectivity of 90.4%.

$$P(f(X) = 0, Y = 0) = P(f(X) = 0|Y = 0)P(Y = 0)$$
$$= (0.904)(0.99) \approx 0.895$$

$$P(f(X) = 1, Y = 0) = P(f(X) = 1|Y = 0)P(Y = 0)$$
$$= (0.096)(0.99) \approx 0.095$$

$$P(f(X) = 0, Y = 1) = P(f(X) = 0|Y = 1)P(Y = 1)$$
$$= (0.2)(0.01) = 0.002$$

$$P(f(X) = 1, Y = 1) = P(f(X) = 1|Y = 1)P(Y = 1)$$
$$= (0.8)(0.01) = 0.008$$

Classified As:

Correct Label:	0	1
0	0.895	0.095
1	0.002	0.008

Quiz

Go toPrairieLearn, try the quiz!

Summary

- Bayes Error Rate:

$$\text{Bayes Error Rate} = \sum_{x} P(X = x) \min_{y} P(Y \neq y | X = x)$$

- Confusion Matrix, Precision & Recall

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}$$