Displays for Statistics 5401

Lecture 34

November 28, 2005

Christopher Bingham, Instructor
612-625-1024

Class Web Page
http://www.stat.umn.edu/~kb/classes/5401

Examples of Classification Problems

Classification is the process of "guessing" on the basis of data **x**, which population, π_1, π_2, ..., or π_g, an individual is in.

Classification can be part of various tasks:

Diagnosis of a medical condition on the basis of a patient's data

- Each *population* π_j consists of individuals with a particular **health condition** from a list of g conditions (one might be "no disease").

  Data **x** are the patient's medical history and results of medical diagnostic procedures carried out on the patient.

Effect of rarity

A physician might be reluctant to diagnose a very rare disease, even if the symptoms were more consistent with it than with other more common conditions.

Predicting bankruptcy on the basis of an individual's credit history

- π_1 is the *population* of individuals seeking credit who will declare bankruptcy in the next 12 months
- π_2 is the population of such people that will *not* do so (g = 2).

Identification of a variety or species using data from an individual organism

- The *populations* π_j are different varieties of a particular species or variety of organism.
- Data **x** are various measured characteristics such as petal length.

In some areas, this is done by a procedure based on a "key" which can be summarized by a "tree" of choice. Methods of this type are CART and FIRM.

Prior probabilities quantify rarity

You quantify the rarity of a population by its *prior probability*,

$$p_i \equiv P(\pi_i).$$

p_i is the probability, ***prior to observing*** data **x**, that a case belongs to or comes from population π_i.

Knowledge of p_1, ..., p_g almost certainly should affect your choice of a classification rule.

When p_i is small, individuals from π_i are rare and you probably should require stronger evidence to classify an individual as belonging to π_i.

When p_i is close to 1, your should require strong evidence to classify an individual as anything other than from π_i.

- For *diagnosis*, p_i measures how *prevalent* medical condition i is among the patients seen by the physician. A rare condition has small p_i.
- For bankruptcy, $p_1 = 1 - p_2 =$ P(randomly selected loan applicant will declare bankruptcy).
- For *identifying* plant varieties, p_i = proportion (prevalence) of plants of variety i in all plants of that type.

  Alternatively, p_i might measure a combination of actual prevalence and ease of finding or collecting specimens the variety. It's possible a common variety is very hard to see.

Because we assume **x** comes from one of g specific populations, $\sum_{1\leq i\leq g} p_i = 1$.

By Bayes' rule, once you know **x**, the **posterior probability** $P(\pi_i \mid \mathbf{x})$ that **x** comes from population π_i is

$$P(\pi_i \mid \mathbf{x}) = \frac{p_i f_i(\mathbf{x})}{\sum_{1\leq j\leq g} p_j f_j(\mathbf{x})}$$

The *numerator* weights the density in π_i by the prior probability of π_i.

- Large p_i can compensate for small $f_i(\mathbf{x})$.
- $P(\pi_i \mid \mathbf{x})$ is large when the prior probability p_i is large *and* $f_i(\mathbf{x})$ is large.

The *denominator* is exactly what is needed so that $\sum_{1\leq j\leq g} P(\pi_j \mid \mathbf{x}) = 1$.

- It is the marginal distribution of **x** when you pick a population using p_i and observe **x** with density $f_i(\mathbf{x})$.

Classification Rules

I use the notation $\hat{\pi}(\mathbf{x})$ as a generic symbol for a procedure, rule or formula to select π on the basis of data $\mathbf{x}$.

- When the procedure selects π_i based on data $\mathbf{x}$, you write $\hat{\pi}(\mathbf{x}) = \pi_i$.
- The possible "values" for $\hat{\pi}(\mathbf{x})$ are π_1, π_2, ..., π_g.

The notation reflects a view of classification as an *estimation* procedure, where the unknown "parameter" is π_i.

Equivalently, i is an unknown parameter, leading to the clumsier notation $\pi_{\hat{i}} = \pi_{\hat{i}(\mathbf{x})}$ where $\hat{i}(\mathbf{x})$ is the index chosen.

A sensible rule: $\hat{\pi}(\mathbf{x})$ = population π_i with largest posterior probability $P(\pi_i | \mathbf{x})$.

How do you compare two rules?

When you can answer this question, you can then ask:

> Which rule, if any, is the **best rule** in the sense of being at least as good as (no worse than) any other rule.

When g = 2, this issue has a lot in common with testing a null hypothesis H_0 (population π_1) against an alternative H_a (population π_2).

There, you want the probabilities of incorrect choices $\alpha \equiv P(\text{reject} | H_0)$ (type I error) and $\beta \equiv P(\text{not reject} | H_a)$ (type II error) to be small. Equivalent you want probabilities $1 - \alpha$ and $1 - \beta$ of correct choices to be large.

This suggests error probabilities are a way to in evaluate classification rules.

(Mis)classification probabilities

Notation

$P(i \mid j) = P(\hat{\pi}(\mathbf{x}) = \pi_i \mid \pi_j)$, $1 \le i,j \le g$

= P(classify **x** as from π_i when it actually is from π_j).

A more complete notation would $P_{\hat{\pi}}(i \mid j)$ or $P(i \mid j; \hat{\pi})$ since $P(i \mid j)$ depends on $\hat{\pi}$.

Trivial example: $\hat{\pi} \equiv$ *Always* choose π_1

$P(1 \mid 1) = 1$; $P(j \mid 1) = 0$, $j \neq 1$

$P(1 \mid \ell) = 1$; $P(j \mid \ell) = 0$, $j \neq 1$, $\ell \neq 1$

Less trivial example with p = 1.

π_1: x is $N(30,5^2)$; π_2: x is $N(40,7^2)$.

Suppose $\hat{\pi}(x)$ selects π_1 when $x \le 35$ and selects π_2 when $x > 35$. Then from

```
Cmd> cumnor(vector((35-30)/5,(35-40)/7))
(1)     0.84134     0.23753     P(x ≤ 35|π1) and P(x ≤ 35|π2)
```

$P(1 \mid 1) = .841$	$P(2 \mid 1) = .159$	1
$P(1 \mid 2) = .238$	$P(2 \mid 2) = .762$	1

What's the overall probability of error?

More about $P(i \mid j)$

- $\sum_{1 \le i \le g} P(i \mid j) = 1$ (always select some π)
- $P(j \mid j) =$ P(**x** from π_j *correctly* classified)
- $1 - P(j \mid j) = \sum_{i \neq j} P(i \mid j) =$ P(**x** from π_j *misclassified*).

You can display $P(i \mid j)$ in a g by g table:

Prior		Classification Decision				
Pop	Pr	π_1	π_2	π_3	...	π_g
π_1	p_1	$P(1 \mid 1)$	$P(2 \mid 1)$	$P(3 \mid 1)$	...	$P(g \mid 1)$
π_2	p_2	$P(1 \mid 2)$	$P(2 \mid 2)$	$P(3 \mid 2)$	...	$P(g \mid 2)$
π_3	p_3	$P(1 \mid 3)$	$P(2 \mid 3)$	$P(3 \mid 3)$	...	$P(g \mid 3)$
...	...	...	...	...	...	...
π_g	p_g	$P(1 \mid g)$	$P(2 \mid g)$	$P(3 \mid g)$	...	$P(g \mid g)$

The off diagonal elements $P(j \mid i)$, $j \neq i$ are generalizations of α and β in a hypothesis test. The diagonal elements are analogous to $1 - \alpha$ and $1 - \beta$.

- The off-diagonal elements are probabilities of *incorrect* classification. They generalize α and β in a hypothesis test
- The diagonal elements $P(j \mid j)$ are probabilities of *correct* classification. They generalize $1 - \alpha$ and $1 - \beta$ in a hypothesis test
- You want $P(i \mid j)$ to be small, $j \neq i$.
- You want $P(i \mid i)$ to be large.

The ***Total Probability of Misclassification*** (**TPM**) of rule $\hat{\pi}$ is

$TPM = TPM(\hat{\pi}) \equiv$

P(misclassify randomly selected case)

Random selection of a case means:

- random select a population π_i with distribution f_i using prior probabilities $p_1, p_2, \ldots p_g$.
- Randomly select **x** from that population

The notation $TPM(\hat{\pi})$ emphasizes that TPM depends on the rule $\hat{\pi}$.

TPM is one answer to the question of how to compare two rules $\hat{\pi}^{(1)}$ and $\hat{\pi}^{(2)}$

$\hat{\pi}^{(2)}$ is better than $\hat{\pi}^{(1)}$ when

$$TPM(\hat{\pi}^{(2)}) < TPM(\hat{\pi}^{(1)})$$

This suggests, a "best" rule would be a rule whose TPM is as small as possible, that is a rule $\hat{\pi}$ that *minimizes* $TPM(\hat{\pi})$.

Formula for TPM($\hat{\pi}$)

When **x** actually comes from π_i the probability it is misclassified by $\hat{\pi}$ is

$$P(\text{misclassify} \mid \mathbf{x} \text{ from } \pi_i) = \sum_{j\neq i} P(j \mid i) = 1 - P(i \mid i)$$

Taking into account the prior probabilities, this means that

$$TPM = TPM(\hat{\pi}) \equiv \sum_{1\leq i\leq g} p_i\{\sum_{j\neq i} P(j \mid i)\}$$
$$= \sum_{1\leq i\leq g} p_i - \sum_{1\leq i\leq g} p_i P(i \mid i) = 1 - \sum_{1\leq i\leq g} p_i P(i \mid i)$$
$$= 1 - P(\text{correct classification})$$

TPM depends explicitly on the prior probabilities p_i.

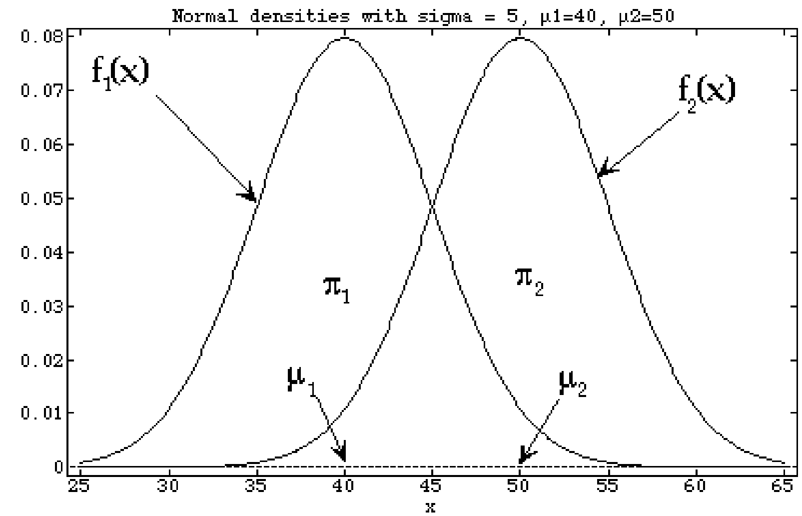
When $\hat{\pi}$ is such that $TPM(\hat{\pi}) \leq TPM(\hat{\pi}')$, for every $\hat{\pi}' \neq \hat{\pi}$, then $\hat{\pi}$ is a **minimum TPM rule**.

Example

Suppose g = 2 and

$$f_1(x) = N(\mu_1,\sigma), \quad f_2(x) = N(\mu_2,\sigma)$$

where $\mu_1 = 40$, $\mu_2 = 50$, $\sigma = 5$



Any sensible rule will be of the form

$$\hat{\pi}_\zeta(x) = \begin{cases} \pi_1 & , x \leq \zeta \\ \pi_2 & , x > \zeta \end{cases}, \text{ some } \zeta$$

That is, there is a single cut point ζ dividing values of x.

$P_\zeta(2 \mid 1) = P(Z > (\zeta - \mu_1)/\sigma)$

$= \texttt{cumnor}((\zeta - \mu_1)/\sigma\texttt{,upper:T})$

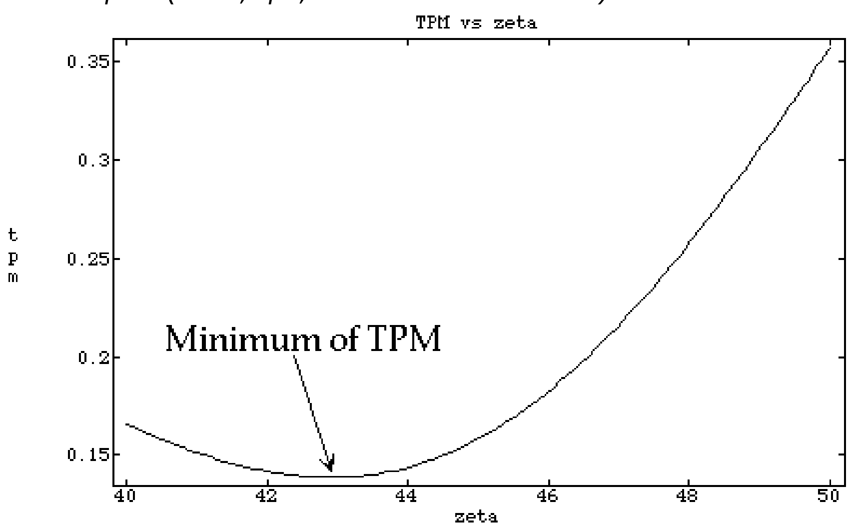
$P_\zeta(1 \mid 2) = P(Z \le (\zeta - \mu_2)/\sigma)$

$= \texttt{cumnor}((\zeta - \mu_2)/\sigma)$

So TPM_ζ =

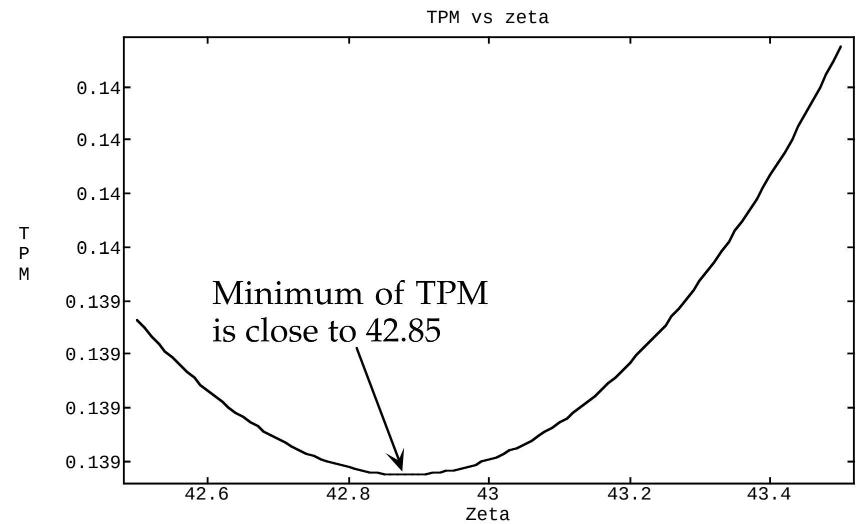
$p_1 P(Z > (\zeta - \mu_1)/\sigma) + p_2 P(Z \le (\zeta - \mu_2)/\sigma)$

```
Cmd> p1 <- .3; p2 <- 1 - p1 # Prior Probabilities
Cmd> mu1 <- 40; mu2 <- 50; sigma <- 5
Cmd> zeta <- run(mu1,mu2,(mu2 - mu1)/100)
Cmd> tpm <- p1*cumnor((zeta - mu1)/sigma,upper:T) + \
          p2 *cumnor((zeta - mu2)/sigma)
Cmd> lineplot(zeta,tpm,title:"TPM vs zeta")
```



Let's focus on values of ζ near where TPM_ζ is minimized, say $42.5 \le \zeta \le 43.5$.

```
Cmd> zeta1 <- 42.5 + run(0,1,.01)
Cmd> tpm1 <- p1*cumnor((zeta1 - mu1)/sigma,upper:T) + \
          p2 *cumnor((zeta1 - mu2)/sigma)
Cmd> lineplot(Zeta:zeta1,TPM:tpm1,title:"TPM vs zeta")
```



The best cutpoint is somewhat nearer μ_2 than to μ_1. This can be expected because the prior probability of π_2 is $p_2 = .7$ as opposed to $p_1 = .3$.

Costs

The <u>probability</u> of misclassification is only one aspect of rule quality.

There are often *costs* or <u>consequences</u> arising from <u>particular</u> misclassifications.

The <u>cost of a misclassification</u> depends on

- The <u>actual</u> population π_i that **x** comes from
- The <u>guessed</u> population $\hat{\pi}(\mathbf{x})$.

Examples:

- The cost of misclassifying an *edible* mushroom as being *poisonous* is certainly less than the cost of *mis*-classifying a *poisonous* mushroom as *edible*.
- The cost of failing to correctly diagnose a hemophiliac (bleeder) about to undergo an operation might be very large (<u>false negative is worse than false positive</u>).

Statistics 5401 Lecture 34 November 28, 2005

Notation

$C(j|i)$ = **cost incurred** when $\hat{\pi}(\mathbf{x}) = \pi_j$ and true population is π_i

It seems reasonable that $C(i|i) \leq 0$, because a negative "cost" is a "benefit".

You can display the values of $C(j|i)$ in a table like that for $P(j|i)$.

Prior		Classification Decision				
Pop	Pr	π_1	π_2	π_3	...	π_g
π_1	p_1	$C(1\|1)$	$C(2\|1)$	$C(3\|1)$	...	$C(g\|1)$
π_2	p_2	$C(1\|2)$	$C(2\|2)$	$C(3\|2)$	...	$C(g\|2)$
π_3	p_3	$C(1\|3)$	$C(2\|3)$	$C(3\|3)$	...	$C(g\|3)$
...	...	...	...	...	...	...
π_g	p_g	$C(1\|g)$	$C(2\|g)$	$C(3\|g)$	...	$C(g\|g)$

Unlike $P(i|j) = P_{\hat{\pi}}(i|j)$, $C(i|j)$ does not depend on $\hat{\pi}$ or $p_1, \ldots, p_g$.

Before you observe **x** from π_i, you don't know the actual cost of classifying it using $\hat{\pi}$, because you don't know how you will classify **x**.

You can, however, find the expected cost of classifying an **x** that comes from π_i. You weight the costs of each possible classification by the probability of that classification:

$$EC(i) = EC_{\hat{\pi}}(i)$$
$$= E[\text{cost} \mid \pi_i] = \sum_{1 \le j \le g} P(j \mid i)C(j \mid i)$$

Now you can use the prior probabilities $\{p_i\}$ to find the overall *expected cost* of classifying a single **x**

$$EC = EC(\hat{\pi}) = \sum_{1 \le i \le g} p_i EC_{\hat{\pi}}(i)$$
$$= \sum_{1 \le i \le g} p_i \{\sum_{1 \le j \le g} P_{\hat{\pi}}(j \mid i)C(j \mid i)\}$$

Note that this is the expected cost of a particular rule $\hat{\pi}$.

EC is another way to compare $\hat{\pi}$'s.

By this criterion, $\hat{\pi}_1$ is "better than" $\hat{\pi}_2$ when $EC(\hat{\pi}_1) < EC(\hat{\pi}_2)$.

Note: EC is not the only reasonable way to use costs in evaluating $\hat{\pi}$.

Alternatives:

- **Maximum expected cost**

$$\max_i EC(i) = \max_i \{\sum_{1 \le j \le g} P(j \mid i)C(j \mid i)\}$$

If you are a pessimist, a good rule might be one that minimizes the maximum expected cost (minimax rule). Of course, if the population for which this cost is maximum is extremely rare, you may greatly increase your expected cost to protect yourself against a rare event.

- **Weighted maximum expected cost**

$$\max_i \{p_i EC(i)\} = \max_i \{p_i \sum_{1 \le j \le g} P(j \mid i)C(j \mid i)\}$$

This downweights the costs of rare events.

Fact (not hard to demonstrate)

Let EP = expected penalty = expected cost when the costs are replaced by the "penalty" $\tilde{C}(j|i) \equiv C(j|i) - C(i|i)$. The penalty satisfies $\tilde{C}(i|i) = 0$. Then for any two rules

$$EP(\hat{\pi}_1) = EP(\hat{\pi}_2) \Leftrightarrow EC(\hat{\pi}_1) = EC(\hat{\pi}_2)$$

$$EP(\hat{\pi}_1) < EP(\hat{\pi}_2) \Leftrightarrow EC(\hat{\pi}_1) < EC(\hat{\pi}_2)$$

In fact

$$EC(\hat{\pi}_1) - EC(\hat{\pi}_2) = EP(\hat{\pi}_1) - EP(\hat{\pi}_2)$$

Thus you get the same ranking of rules by EP as by EC. This means there you lose no generality, by assuming that $C(i|i) = 0$, $i = 1,\ldots,g$ for which costs $EP(\hat{\pi}) = EC(\hat{\pi})$.

From now on I will assume $C(i|i) = 0$ and will, use ECM = the ***Expected Cost of Misclassification*** in place of EC.

$$ECM(\hat{\pi}) = ECM = \sum_i p_i\{\sum_{j\neq i} P(j|i)C(j|i)\}$$
$$= \sum_i p_i ECM(i)$$

ECM is a weighted average of the expected costs of misclassifying an individual from a population.

Suppose *all misclassification costs are the same*, so that $C(j|i) = c$, $i \neq j$ and $C(i|i) = 0$. Then also

$$ECM(i) = c(1 - P(i|i))$$

and

$$ECM(\hat{\pi}) = c\times\sum_{1\le i\le g} p_i(1 - P(i|i)) = c\times TPM(\hat{\pi}).$$

In this equal cost case, ranking rules by ECM is the same as ranking rules by TPM.

When you can identify differential costs of misclassification, ECM is a way to *rank classification rules*.

You should *prefer* $\hat{\pi}_a$ to $\hat{\pi}_b$ when

$$ECM(\hat{\pi}_a) < ECM(\hat{\pi}_b)$$

Using this approach, the "best" rule is a ***minimum ECM*** rule, that is a rule that with the smallest possible ECM.

When you cannot reasonably specify costs, it's sometimes appropriate to act as if all costs are the same.

In this case, you should rank rules by their overall error rate (TPM).

$\hat{\pi}_a$ is "better" than $\hat{\pi}_b$ when

$$TPM(\hat{\pi}_a) < TPM(\hat{\pi}_b)$$

The "best" rule is the ***minimum TPM*** rule which has the smallest possible TPM.

- The **minimum ECM rule** is a classification rule whose **ECM** is not greater ($\le$) than the ECM of any other rule.
- The **minimum TPM rule** is a classification rule whose **TPM** is not greater ($\le$) than the ECM of any other rule.

Suppose you know how to find a minimum ECM rule for any costs $C(j|i)$.

Then, because TPM = ECM when $C(j|i) = 1$, $j \neq i$, you also know how to determine the minimum TPM rule.