Maximizing the Usefulness of Statistical Classifiers for Two Populations

Daniel R. Jeske
Department of Statistics
University of California – Riverside

Outline

1. Background and Motivation
2. Neutral Zones
3. Two-Stage Procedures
4. Summary

Conference on Predictive Inference and Its Applications
May 7-8, 2018
Iowa State University, Ames

Deaths in the United States

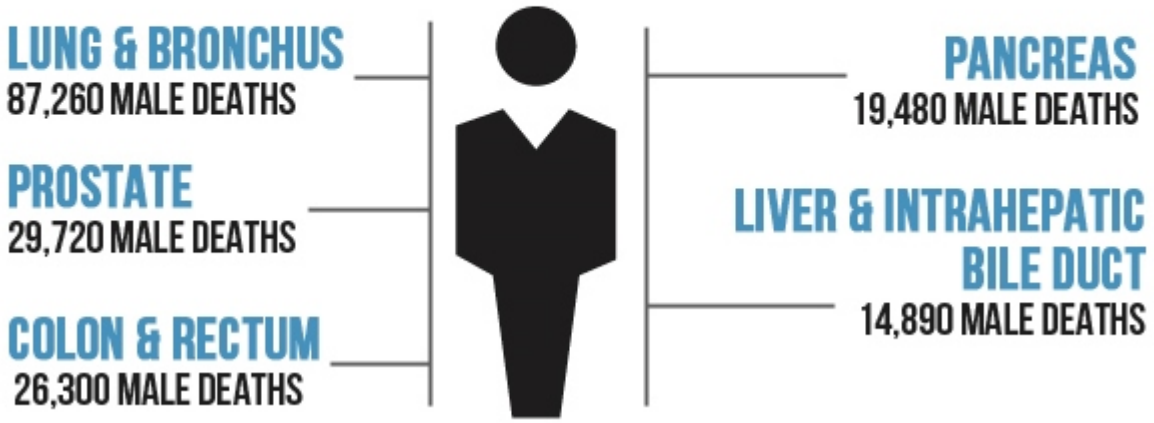
Basic Statistics

Leading Causes of Death (2014 data)

- 2,626,418 deaths
  - Heart disease: 614,348
  - **Cancer: 591,699**
  - Chronic lower respiratory diseases: 147,101
  - Accidents: 136,053
  - Stroke: 133,103
  - Alzheimer's disease: 93,541
  - Diabetes: 76,488
  - Flu and Pneumonia: 55,227
  - Kidney disease: 48,146
  - Suicide: 42,773
  - Homicide: 14,249 (note: homicide is not the next one in line)

http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

FIVE MOST DANGEROUS CANCERS IN MEN



https://www.unitypoint.org/

Classification of Prostate Cancer Patients

Suppose a patient is considering radical prostatectomy surgery:

- Aggressive prostate cancer, the kind that has already moved (e.g., to bone marrow) will likely recur within 2.5 years of the surgery.

- If we could know recurrence was likely for the patient, then it could be inferred the disease is aggressive, and surgery should complemented with a parallel treatment protocol.

- If we could know recurrence was unlikely It could be inferred the disease is not aggressive and the patient could conceivably opt to delay surgery and start active surveillance.

Classification of Prostate Cancer Patients

Suppose a patient is considering radical prostatectomy surgery:

- Aggressive prostate cancer, the kind that has already moved (e.g., to bone marrow) will likely recur within 2.5 years of the surgery.

- If we could know recurrence was likely for the patient, then it could be inferred the disease is aggressive, and surgery should complemented with a parallel treatment protocol.

- If we could know recurrence was unlikely It could be inferred the disease is not aggressive and the patient could conceivably opt to delay surgery and start active surveillance.

Can a classifier be used prior to surgery to *predict* recurrence within 2.5 years following surgery?

Training Data Set[1]

Data set: 450 patients that had prostate surgery. Recurrence within 2.5 years divides the patients into two groups:

"Recurrence = YES" (56 patients)

"Recurrence = NO" (394 patients)

Before the surgery, each patient was measured with respect to four key variables believed to be indicators of how aggressive the prostate cancer is:

1.Prostate Specific Antigen (PSA)

2.Biopsy Gleason Score (GS)

3.PSA Peptidase Activity (PPA)

4. Prostate Cancer Antigen 3 (PCA3)

[1]Collaboration with Dr. Steven Smith, City of Hope National Medical Center, Duarte, CA

Logistic Regression Fit

Variable	Type of Variable	Estimate	Std. Error	P-value
Intercept		-2.8	.274	< .0001
Gleason Score (GS)	CLIN	0.596	.144	< .0001
PSA	CLIN	1.02	.187	< .0001
PCA3	LAB	0.318	.145	.0282
PPA	LAB	-2.094	.715	.0034

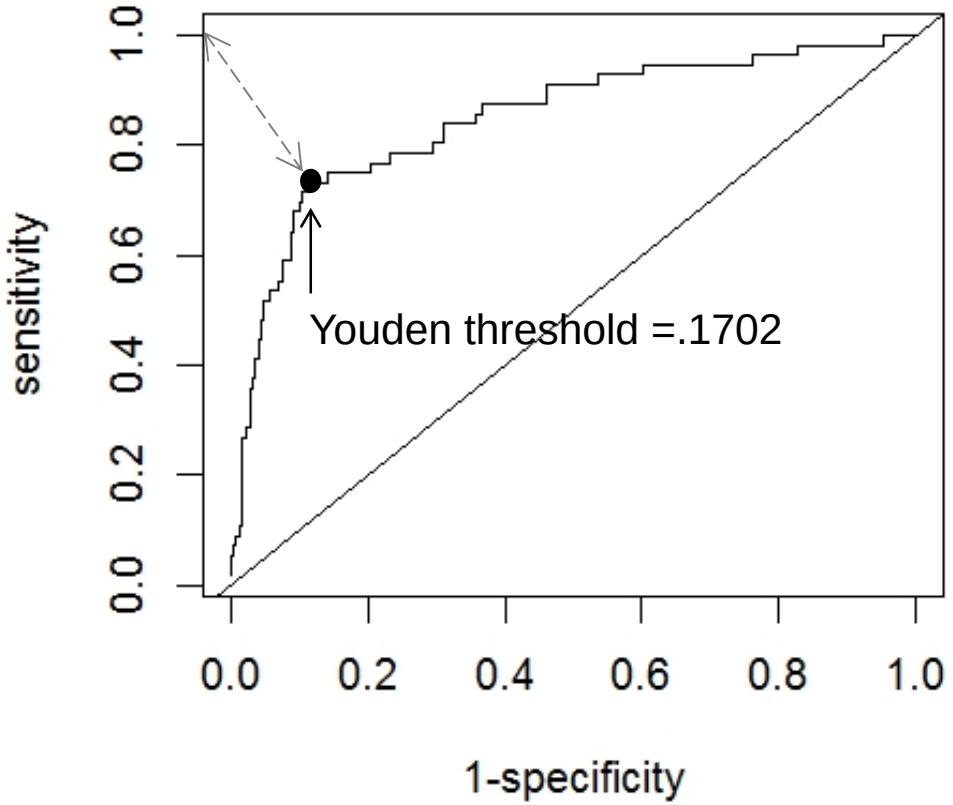
All covariates were centered and standardized using means and standard deviations calculated from the training data

$T = P(\text{Recurrence} \mid \text{GS, PSA, PCA3, PPA})$

$$= \frac{\exp(-2.8 + .596 \times \text{GS} + 1.02 \times \text{PSA} + 0.318 \times \text{PCA3} - 2.094 \times \text{PPA})}{1 + \exp(-2.8 + .596 \times \text{GS} + 1.02 \times \text{PSA} + 0.318 \times \text{PCA3} - 2.094 \times \text{PPA})}$$

Logistic Regression Classifier

Decision Statistic: Pr(Recurrence | GS, PSA, PCA3, PPA)



$\hat{C}$

C	0	1
0	88.6%	11.4%
1	26.8%	73.2%

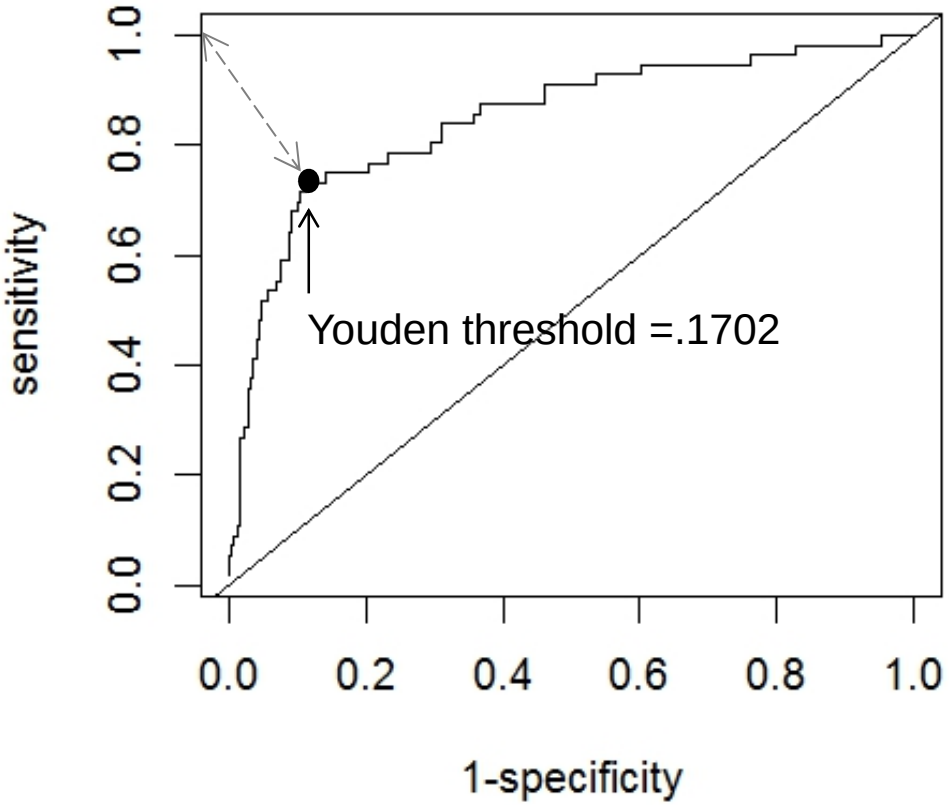
Prevalence $= 12.4\%$

$\text{PPV} = \text{P}(C = 1 \mid \hat{C} = 1) = 47.7\%$

$\text{NPV} = \text{P}(C = 0 \mid \hat{C} = 0) = 95.9\%$

Logistic Regression Classifier

Decision Statistic: Pr(Recurrence | GS, PSA, PCA3, PPA)



$\hat{C}$

C	0	1
0	88.6%	11.4%
1	26.8%	73.2%

Prevalence $= 12.4\%$

$\text{PPV} = \text{P}(C = 1 \mid \hat{C} = 1) = 47.7\%$

$\text{NPV} = \text{P}(C = 0 \mid \hat{C} = 0) = 95.9\%$

Are the relatively high FPR and FNR acceptable? **What could be done about it?**

About Known Unknowns…..

"There are known knowns. These are things we know that we know.

There are known unknowns. That is to say, there are things that we know we don't know.

But there are also unknown unknowns. There are things we don't know we don't know.“

About Known Unknowns…..

"There are known knowns. These are things we know that we know.

There are known unknowns. That is to say, there are things that we know we don't know.

But there are also unknown unknowns. There are things we don't know we don't know.“

Donald Rumsfeld, February 12, 2002
U.S. Secretary of Defense under Pres. George W. Bush

About Known Unknowns…..

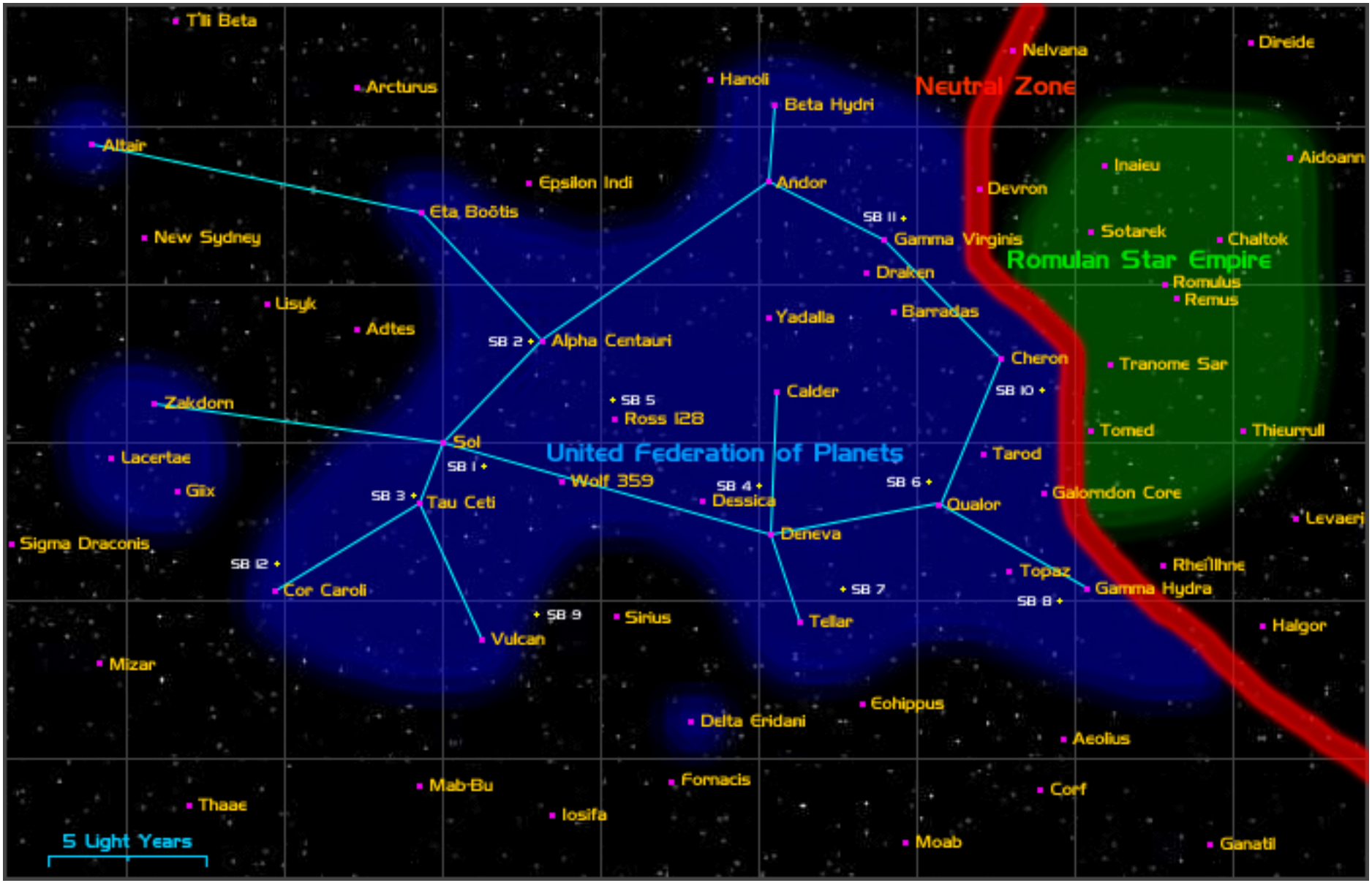
"There are known knowns. These are things we know that we know.

There are known unknowns. That is to say, there are things that we know we don't know.

But there are also unknown unknowns. There are things we don't know we don't know.“

Donald Rumsfeld, February 12, 2002
U.S. Secretary of Defense under Pres. George W. Bush

Neutral Zones



The Earth-Romulan War, was a major interstellar conflict fought from 2156 to 2160 between the forces of United Earth and those of the Romulan Star Empire.

Minimum Cost Neutral Zone Classifier

True Class Label	Predicted Class Label		
	0	1	N
0	0	C_{10}	C_N
1	C_{01}	0	C_N

$$p_1(x) = P(\text{class } 1 \mid x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_0 f_0(x)}$$

$$\hat{C}_B(x) = \begin{cases} 0 & \text{if} \quad p_1(x) \le \dfrac{C_N}{C_{01}} \\ 1 & \text{if} \quad p_1(x) > 1 - \dfrac{C_N}{C_{10}} \\ N & \text{if} \quad \text{otherwise} \end{cases}$$

Neutral zone exists when cost of not making a decision is lower than :

1) posterior expected loss of classifying as 0
2) posterior expected loss of classifying as 1

Related Literature: Rao(1947), Habbema et al. (1974), Swartz (2002, PhD Dissertation), Jeske et al. (2007, 2011, 2014)

Alternative: Fixed Conditional Misclassification Rates

Let T be the decision statistic and assume that large values point toward $C = 1$.

$$\hat{C} = \begin{cases} 1 & \text{if} \quad T \geq C_1 \\ N & \text{if} \quad C_0 < T < C_1 \\ 0 & \text{if} \quad T \leq C_0 \end{cases}$$

Choose the constants to satisfy

$P(\hat{C} = 1 | C = 0) = \alpha$, control of False Positive Rate (FPR)

$P(\hat{C} = 0 | C = 1) = \beta$, control of False Negative Rate (FNR)

Related Literature: Chow (1957), Anderson (1969), Patterson (2016)

Alternative: Fixed Conditional Misclassification Rates

Let T be the decision statistic and assume that large values point toward $C = 1$.

$$\hat{C} = \begin{cases} 1 & \text{if} \quad T \geq C_1 \\ N & \text{if} \quad C_0 < T < C_1 \\ 0 & \text{if} \quad T \leq C_0 \end{cases}$$

Choose the constants to satisfy

$P(\hat{C} = 1 | C = 0) = \alpha$, control of False Positive Rate (FPR)

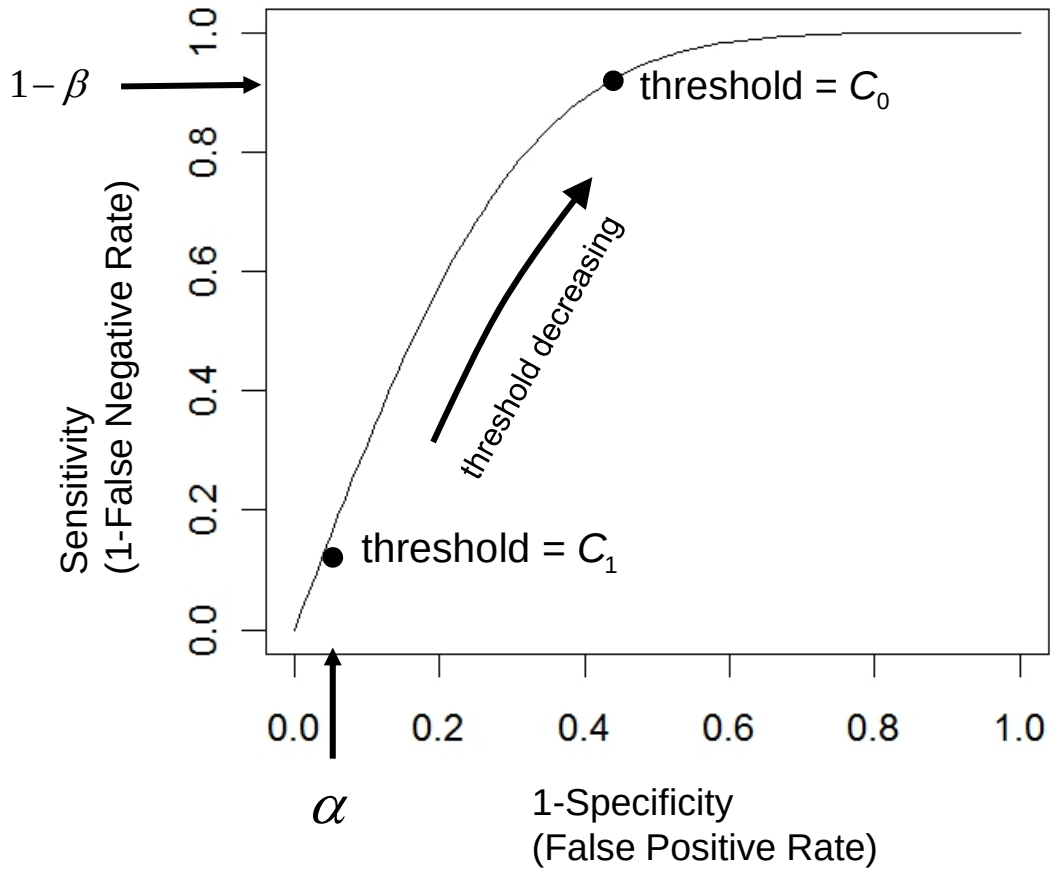
$P(\hat{C} = 0 | C = 1) = \beta$, control of False Negative Rate (FNR)

Key to solving the equations is identifying the conditional distributions of T, given $C = 0$ and $C = 1$. Denote these distributions by F and G, respectively.

Then $C_0 = G^{-1}(\beta)$ and $C_1 = F^{-1}(1-\alpha)$

Construction From ROC Curves

Choose **TWO** thresholds

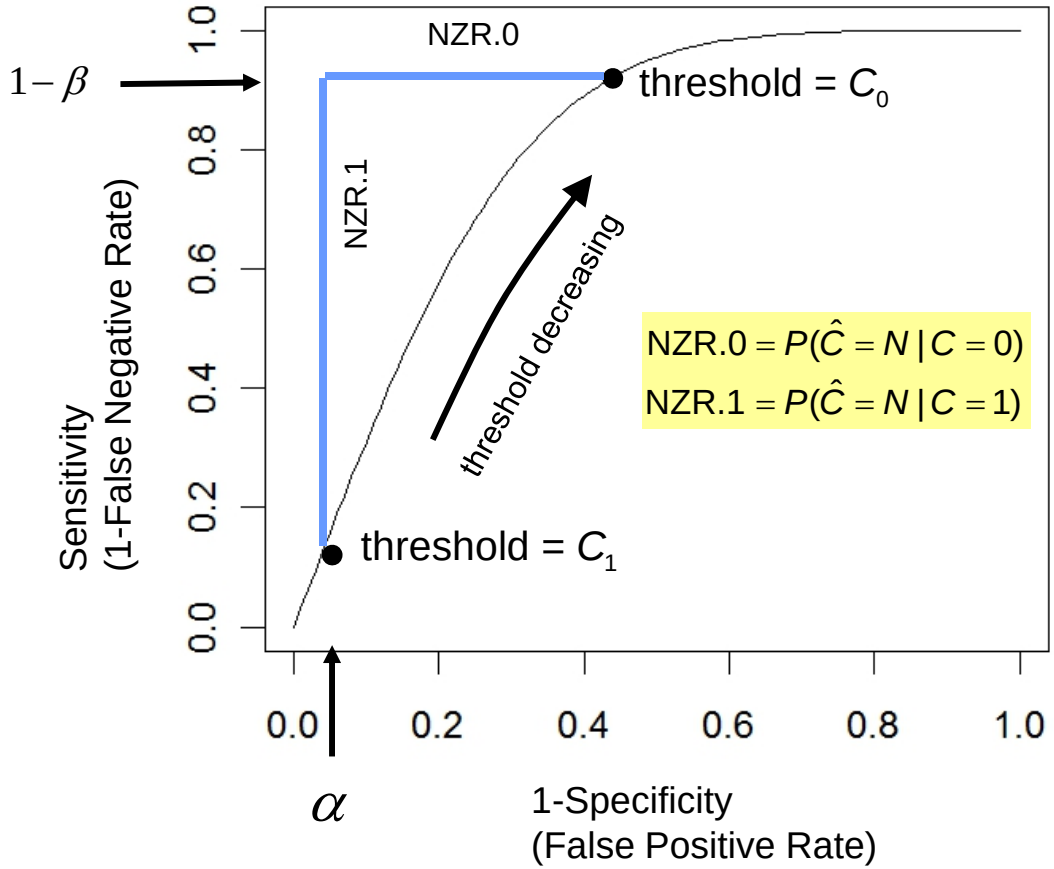


$$\hat{C} = \begin{cases} 1 & \text{if } T \geq C_1 \\ N & \text{if } C_0 < T < C_1 \\ 0 & \text{if } T \leq C_0 \end{cases}$$

The classifier $\hat{C}$ has FPR equal to α and FNR equal to β

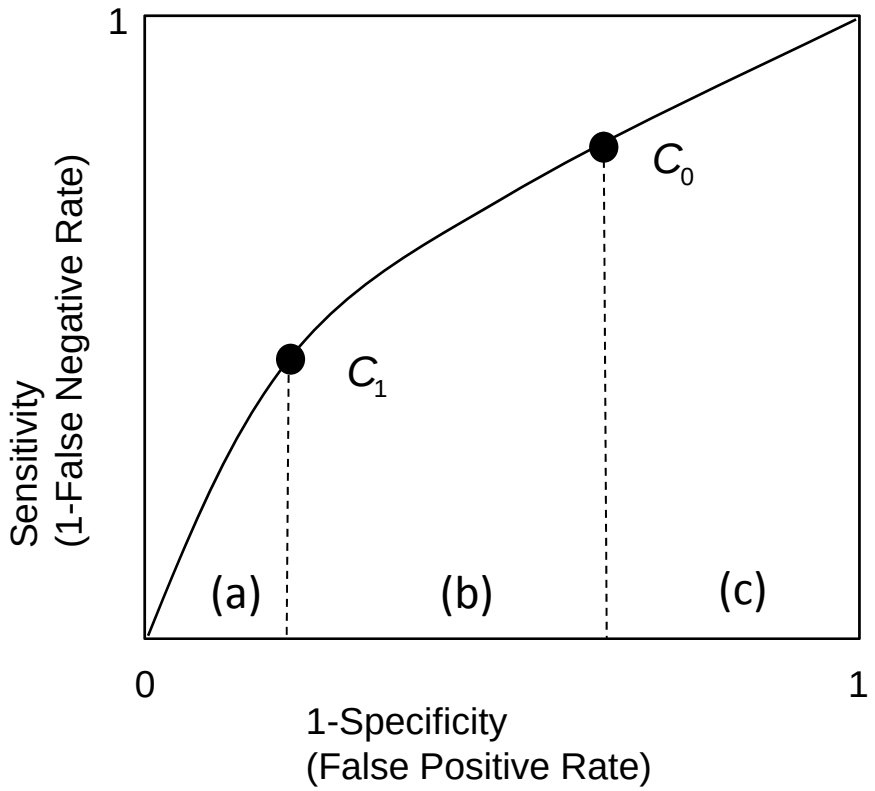
Construction From ROC Curves

Neutral Zone Probabilities



Construction From ROC Curves

Partition of AUC



$T_0 = T$ from a subject from $C = 0$

$T_1 = T$ from a subject from $C = 1$

Well known result: $\text{AUC} = P[T_1 > T_0]$

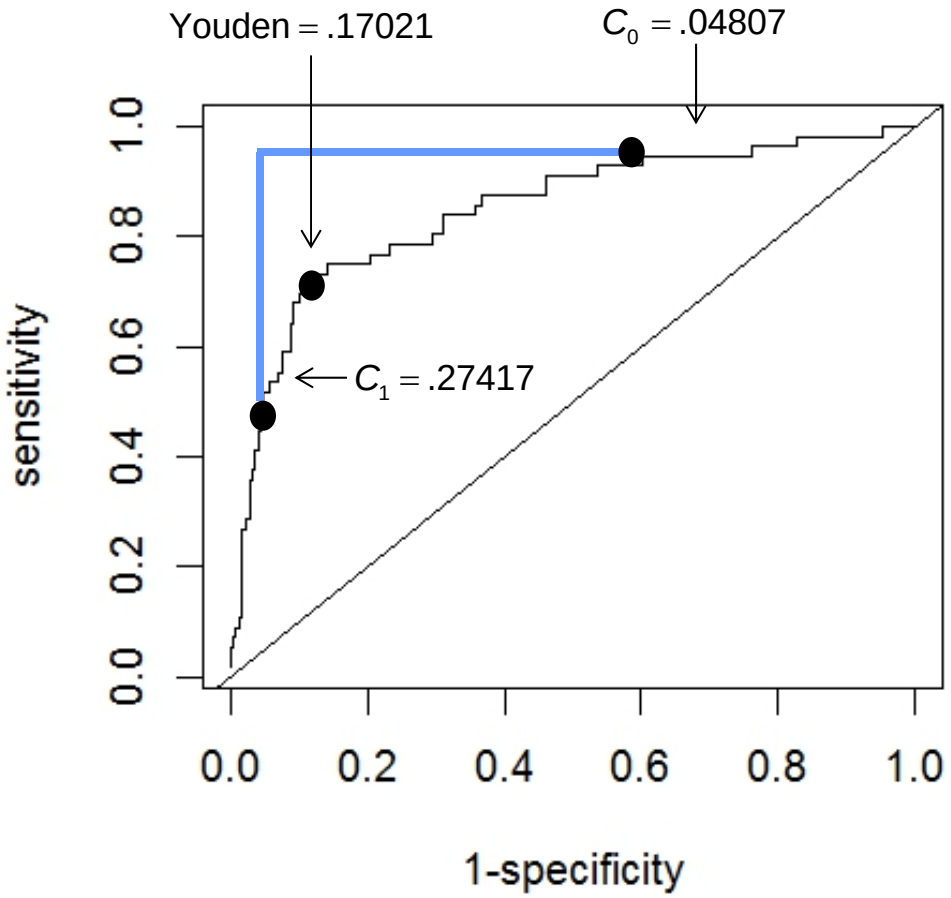
For neutral zone classifier:

$\text{pAUC}_a = P[T_1 > T_0, C_1 < T_0]$

$\text{pAUC}_b = P[T_1 > T_0, C_0 < T_0 < C_1]$

$\text{pAUC}_c = P[T_1 > T_0, T_0 < C_0]$.

Returning to the Prostate Cancer Data



Traditional classifier

C \ $\hat{C}$	0	1
0	88.6%	11.4%
1	28.6%	71.4%

Neutral zone classifier

C \ $\hat{C}$	0	1	N
0	39.9%	4.8%	55.3%
1	5.4%	42.9%	51.7%

Almost half of the patients can receive a reliable assessment, with just a 5% chance of a mistaken judgment. The rest of the patients need some follow-up because their values on PSA, GS, PCA3, and PPA are ambiguous.

PPV and NPV of Neutral Zone Classifier

Proposition

Implementing a neutral zone elevates the PPV and NPV if:

i) $h_G(u) \leq h_F(u)$ (hazard function condition)

ii) $\bar{h}_G(u) \geq \bar{h}_F(u)$ (reversed hazard function condition)

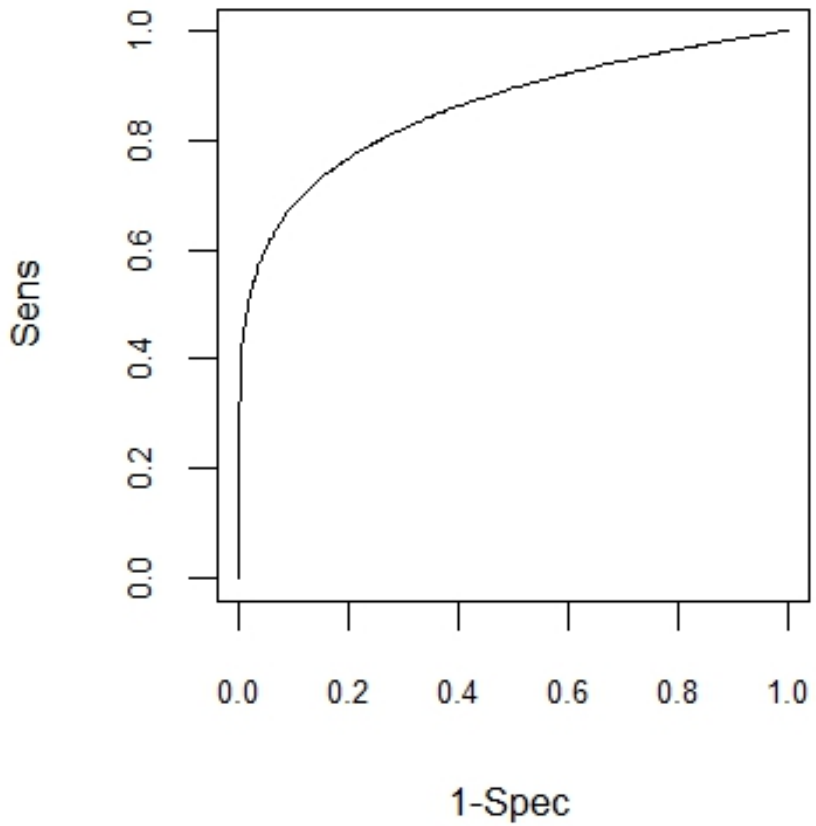
Definitions

hazard function: $h_F(u) = \dfrac{f(u)}{1 - F(u)}$

"reversed" hazard function: $\bar{h}_F(u) = \dfrac{f(u)}{F(u)}$

Returning to the Prostate Cancer Data

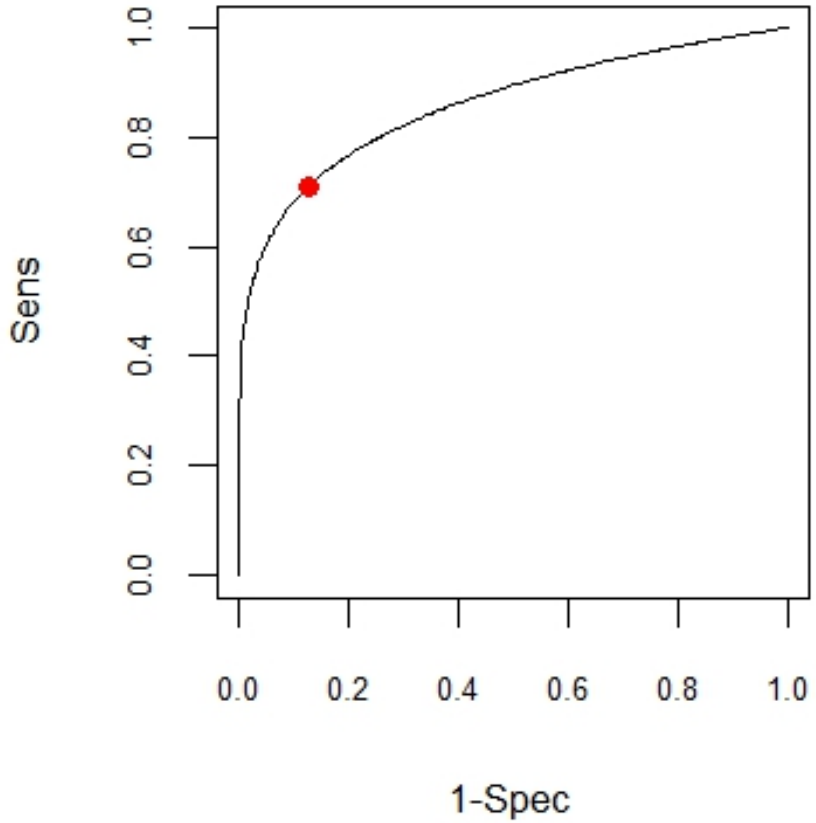
A Useful 4-Way Plot



1. Smoothed ROC Curve

Returning to the Prostate Cancer Data

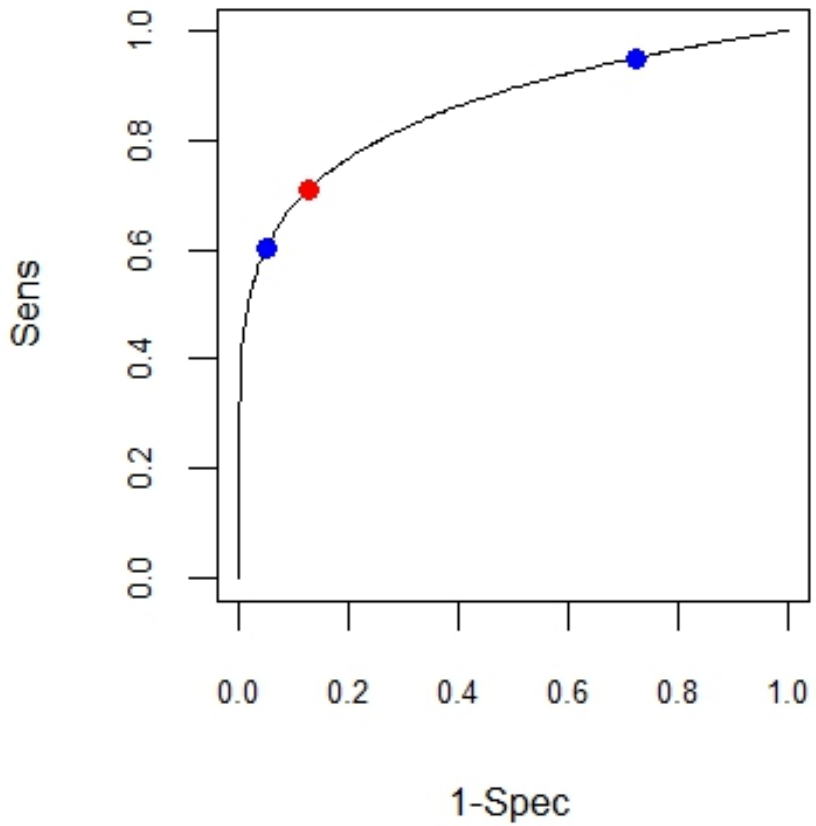
A Useful 4-Way Plot



1. Smoothed ROC Curve
2. Traditional threshold

Returning to the Prostate Cancer Data

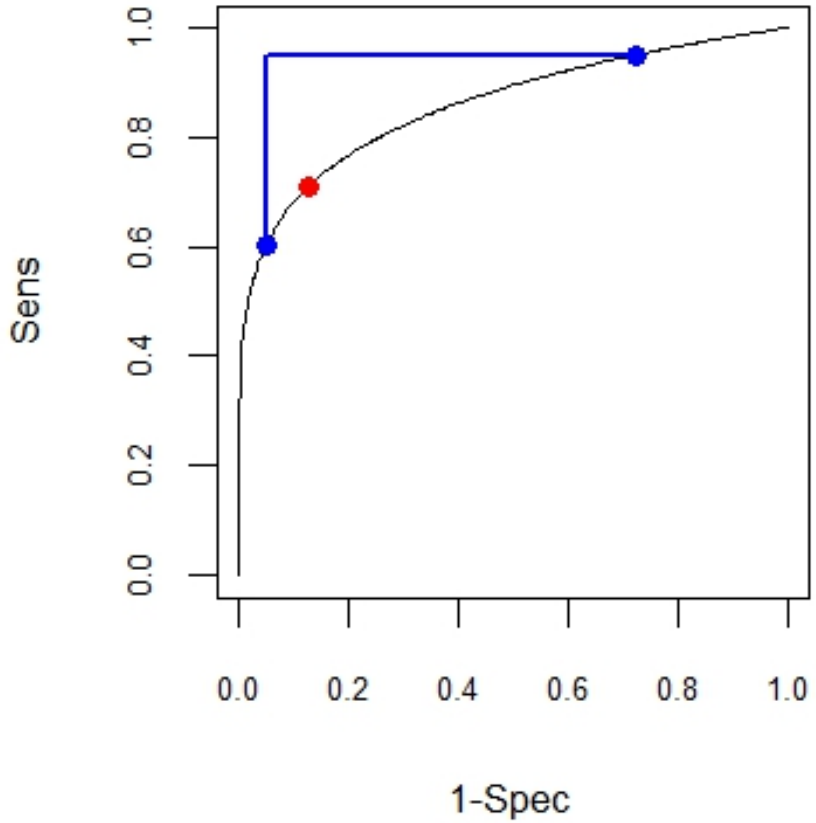
A Useful 4-Way Plot



1. Smoothed ROC Curve
2. Traditional threshold
3. Neutral zone classifier thresholds

Returning to the Prostate Cancer Data

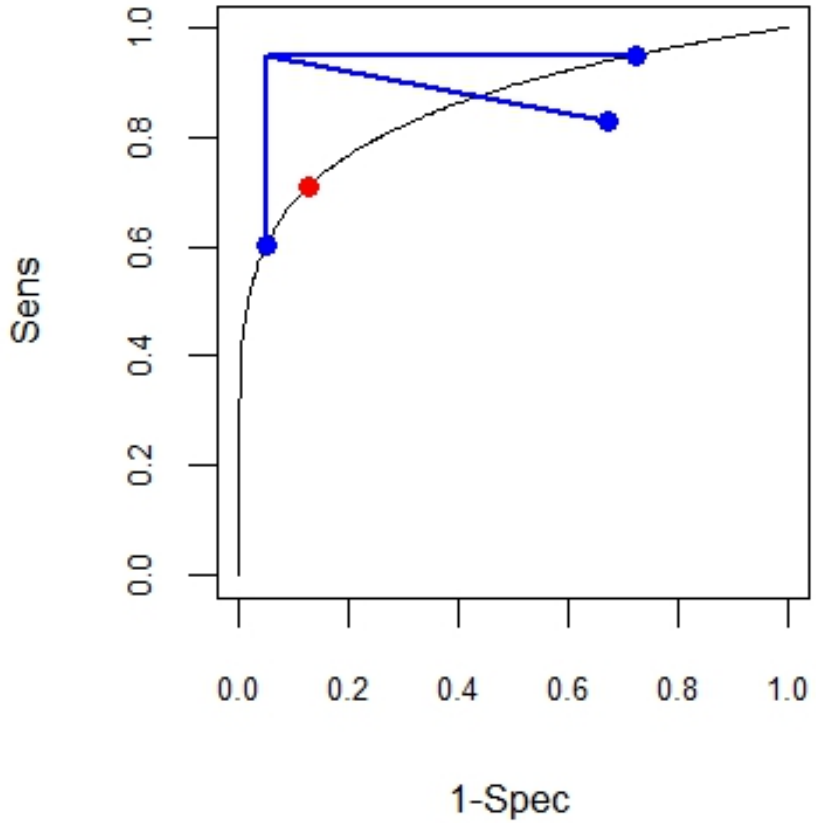
A Useful 4-Way Plot



1. Smoothed ROC Curve
2. Traditional threshold
3. Neutral zone classifier thresholds
4. Visualization of conditional neutral zone probabilities

Returning to the Prostate Cancer Data

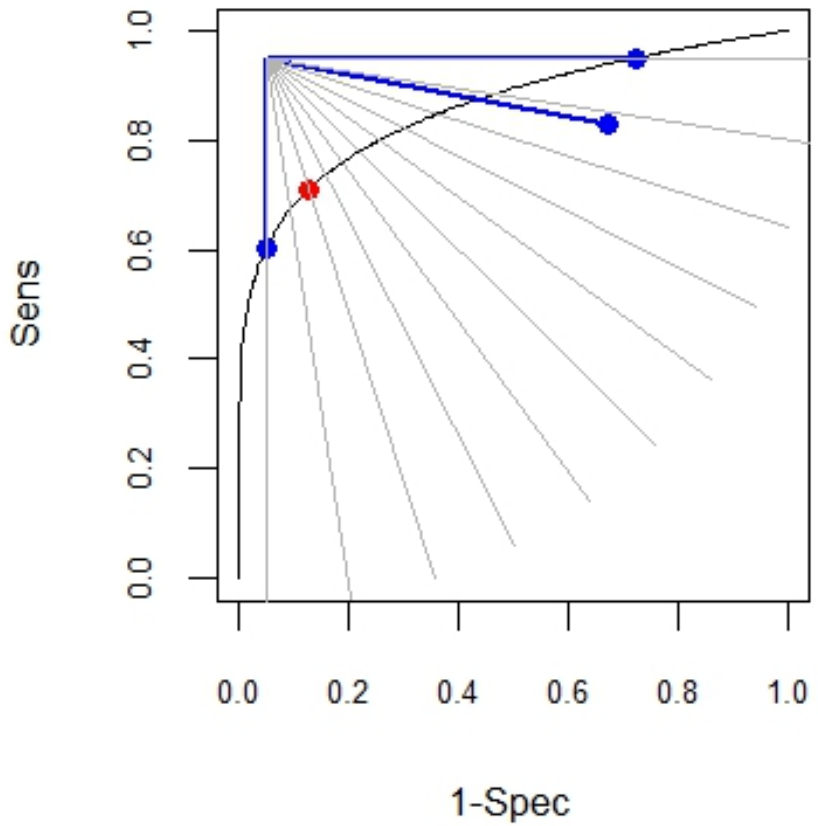
A Useful 4-Way Plot



1. Smoothed ROC Curve
2. Traditional threshold
3. Neutral zone classifier thresholds
4. Visualization of conditional neutral zone probabilities
5. Visualization of unconditional neutral zone probability

Returning to the Prostate Cancer Data

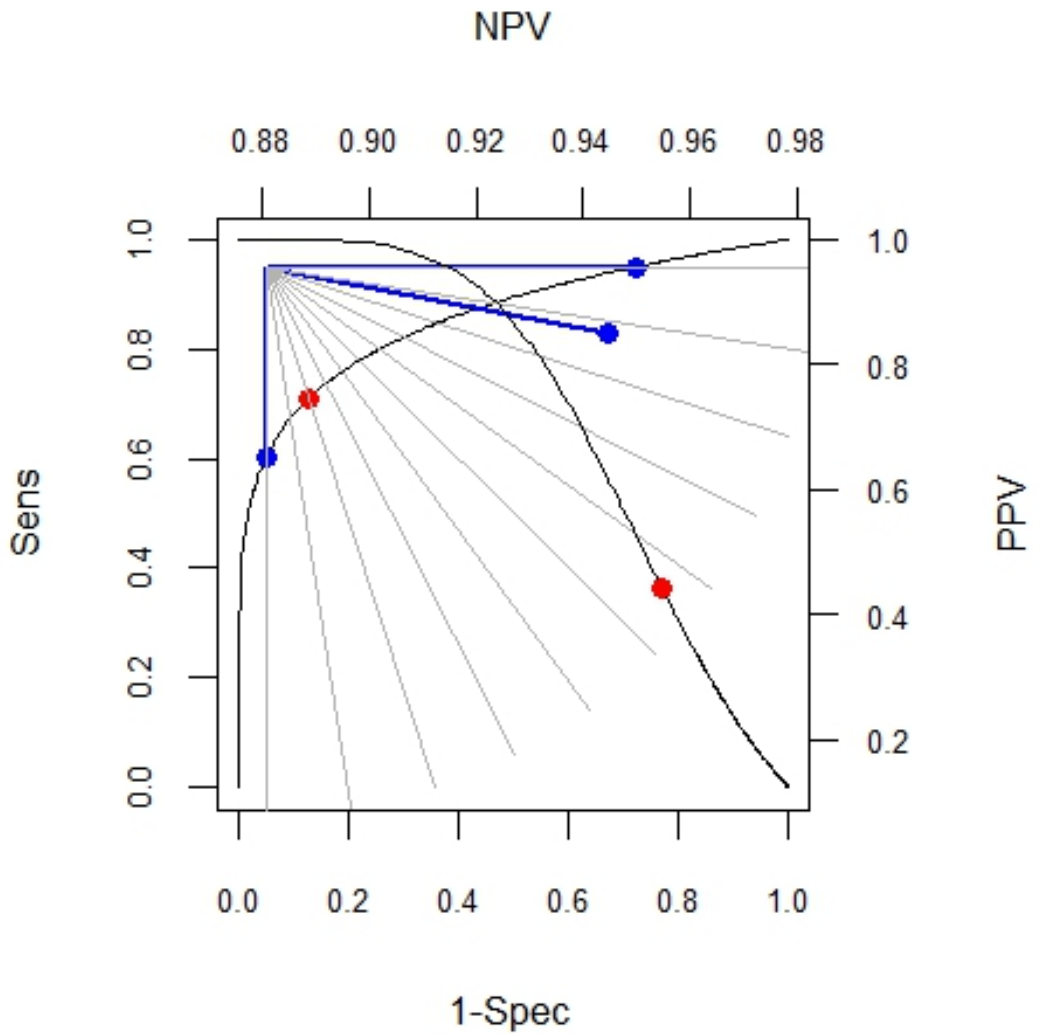
A Useful 4-Way Plot



1. Smoothed ROC Curve
2. Traditional threshold
3. Neutral zone classifier thresholds
4. Visualization of conditional neutral zone probabilities
5. Visualization of unconditional neutral zone probability
6. Unit length reference lines

Returning to the Prostate Cancer Data

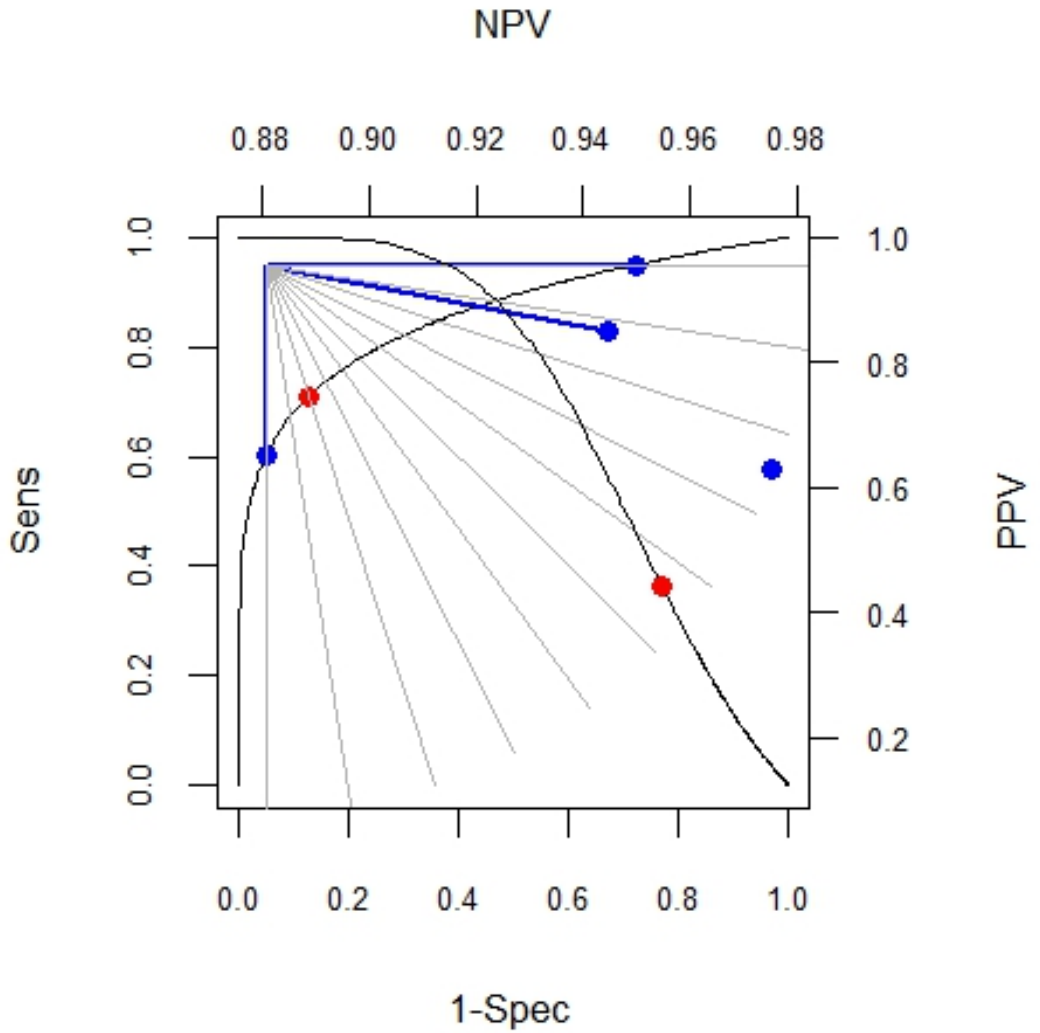
A Useful 4-Way Plot



1. Smoothed ROC Curve
2. Traditional threshold
3. Neutral zone classifier thresholds
4. Visualization of conditional neutral zone probabilities
5. Visualization of unconditional neutral zone probability
6. Unit length reference lines
7. NPV-PPV curve and depiction of values for traditional classifier

Returning to the Prostate Cancer Data

A Useful 4-Way Plot



1. Smoothed ROC Curve
2. Traditional threshold
3. Neutral zone classifier thresholds
4. Visualization of conditional neutral zone probabilities
5. Visualization of unconditional neutral zone probability
6. Unit length reference lines
7. NPV-PPV curve and depiction of values for traditional classifier
8. Depiction of NPV-PPV for traditional classifier

Neutral zones in a Two-Stage Framework

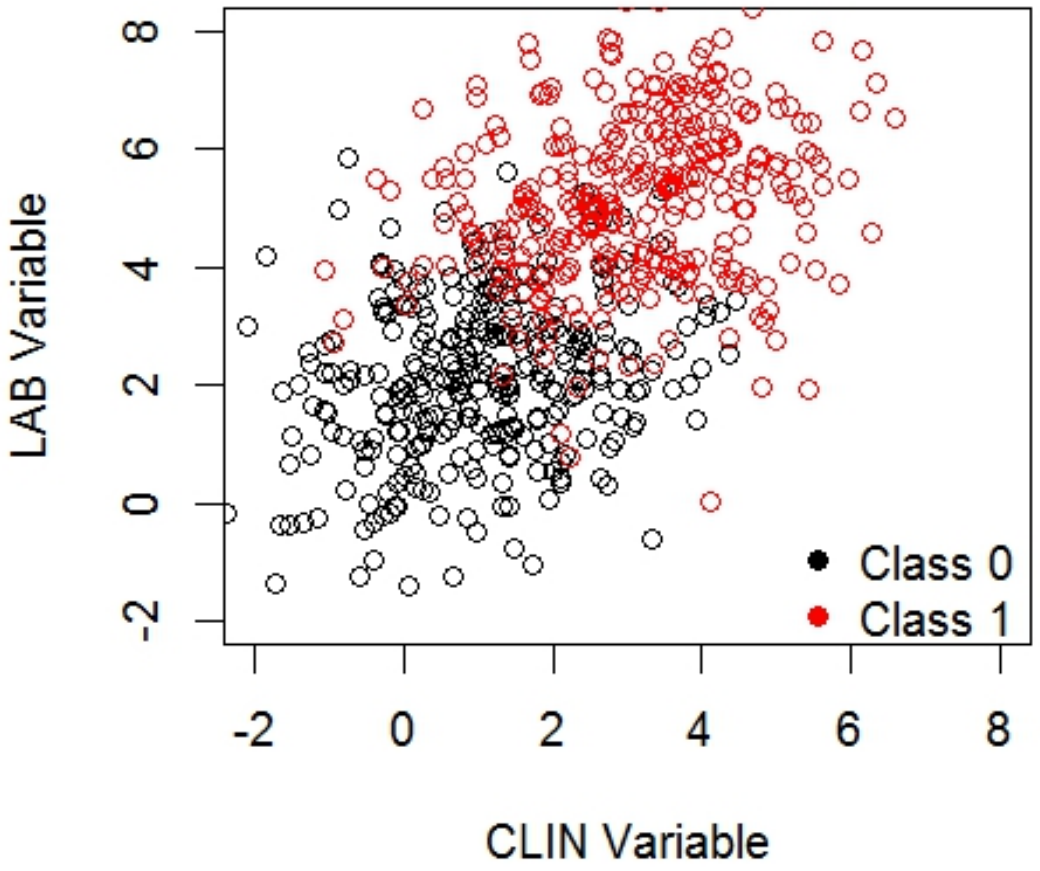
Motivation

- The PCA3 and PPA LAB variables for Prostate Cancer application cost approximately $760 and $1000, respectively.
- Can we develop a useable classifier that gives patients an option to avoid using the expensive LAB variables?
- First stage classifier
  - Utilize only CLIN variables and a neutral zone classifier to identify patients for whom we can predict recurrence reliably
  - Set FPR and FNR to nominally low values, for example 5%, to minimize misclassifications at the first stage.
  - Patients falling into the first stage neutral zone are referred to get the LAB variables for a more conclusive “second stage” decision.

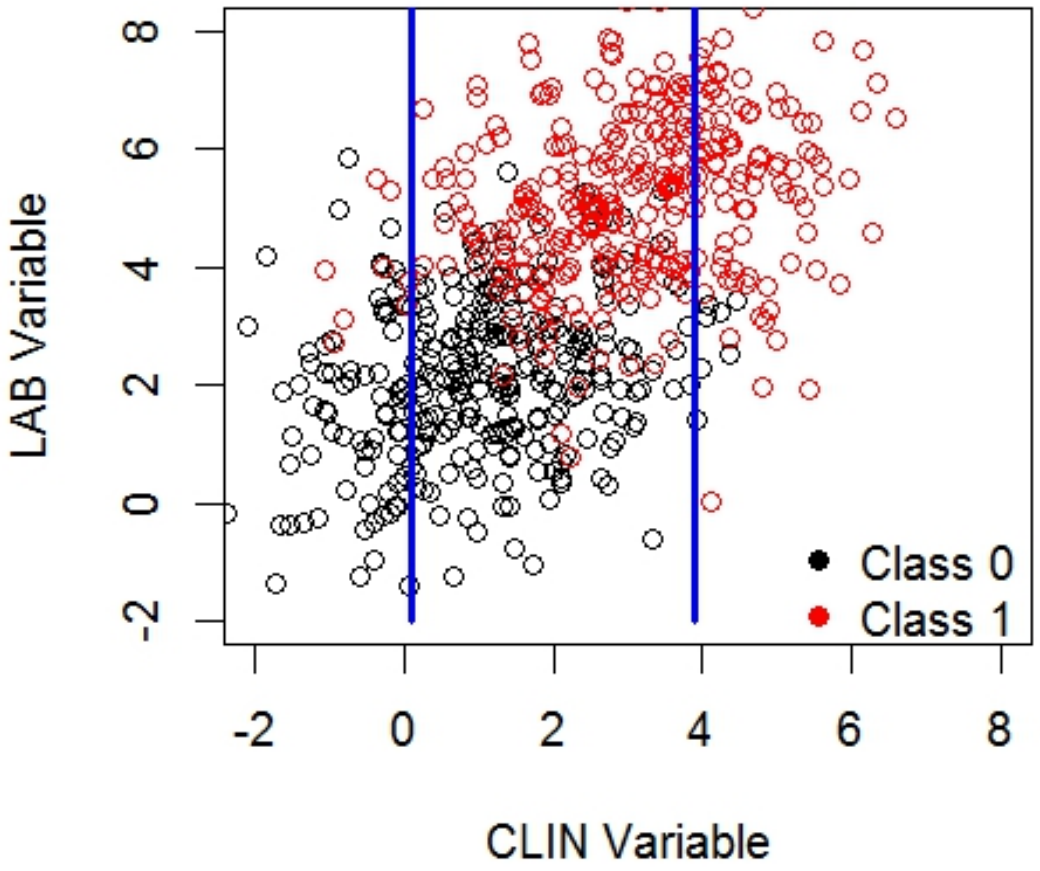
Neutral zones in a Two-Stage Framework

- Second Stage Classifier
  - Using the patients in the first stage neutral zone classifier, fit a second classifier using the CLIN plus the LAB variables.
  - Select a single threshold to ‘force’ a decision at the second stage.

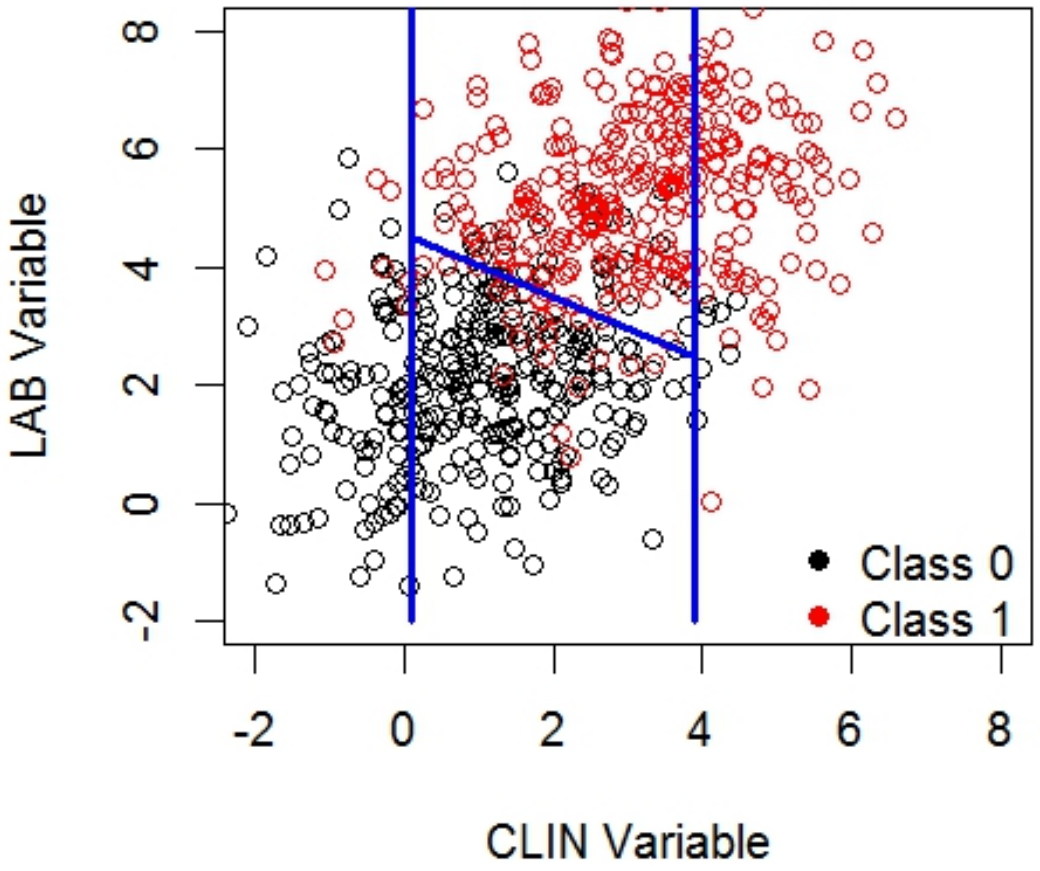
Intuition



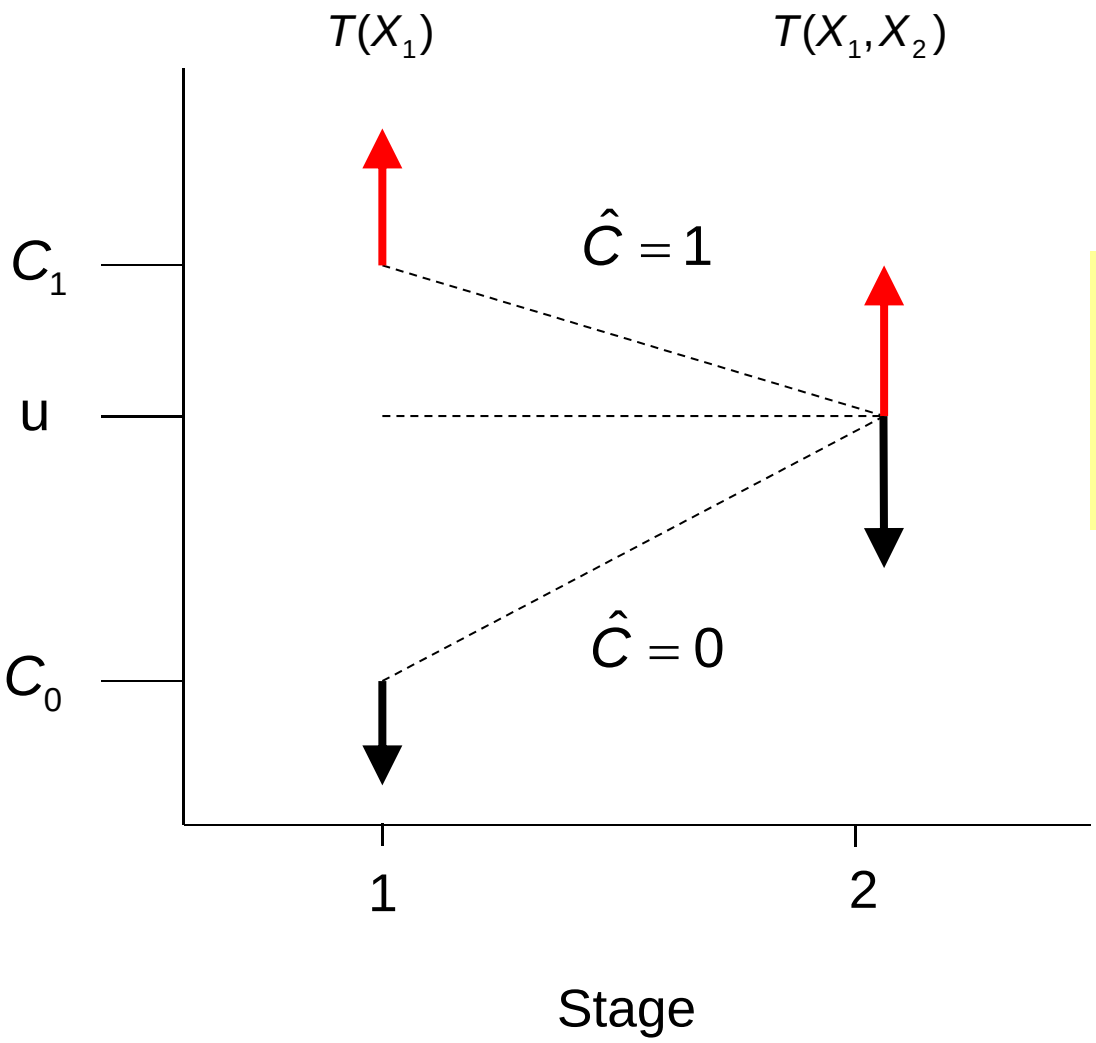
Stage 1 Classifier Uses Only CLIN Variable



Stage 2 Classifier Uses CLIN and LAB Data To Classify Patients Who Fell Into the Neutral Zone



Two-Stage Classifier

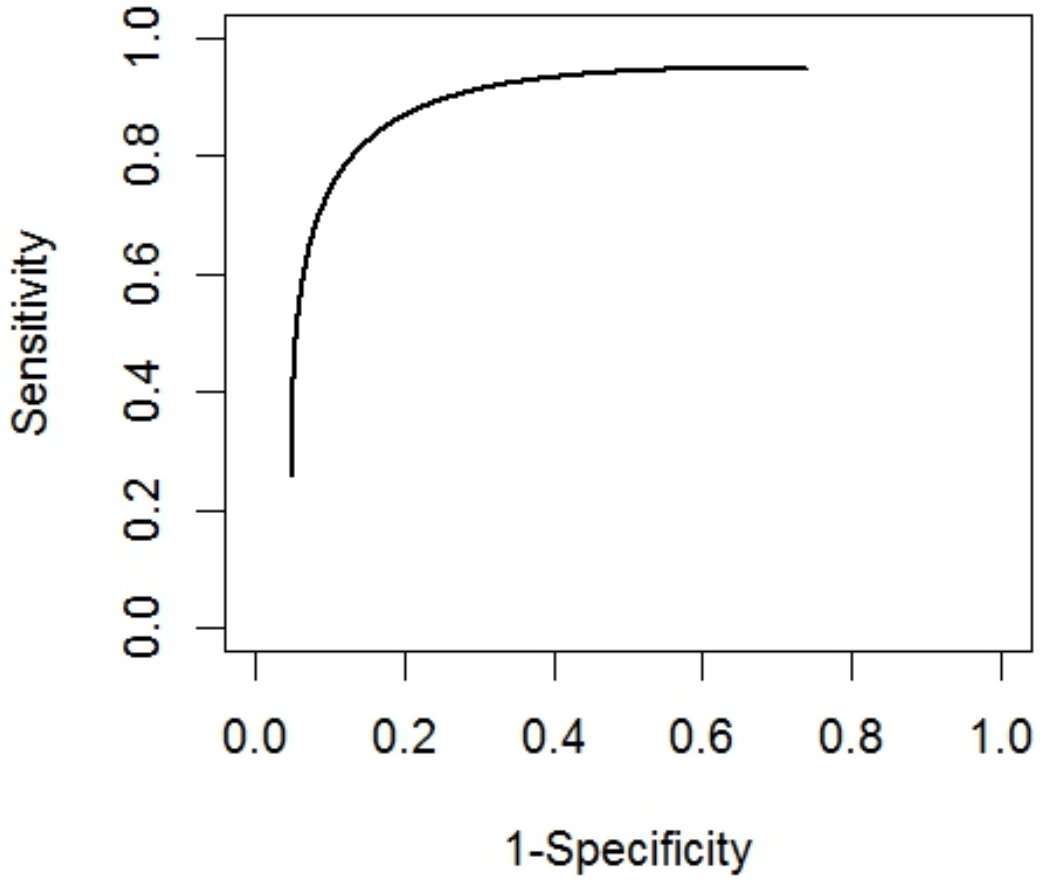


Risk at Stage 1

1) $\Pr\left(T(X_1) \geq C_1 \,|\, C = 0\right) = \alpha$

2) $\Pr\left(T(X_1) \leq C_0 \,|\, C = 1\right) = \beta$

Example ROC Curve of a Two-Stage Classifier

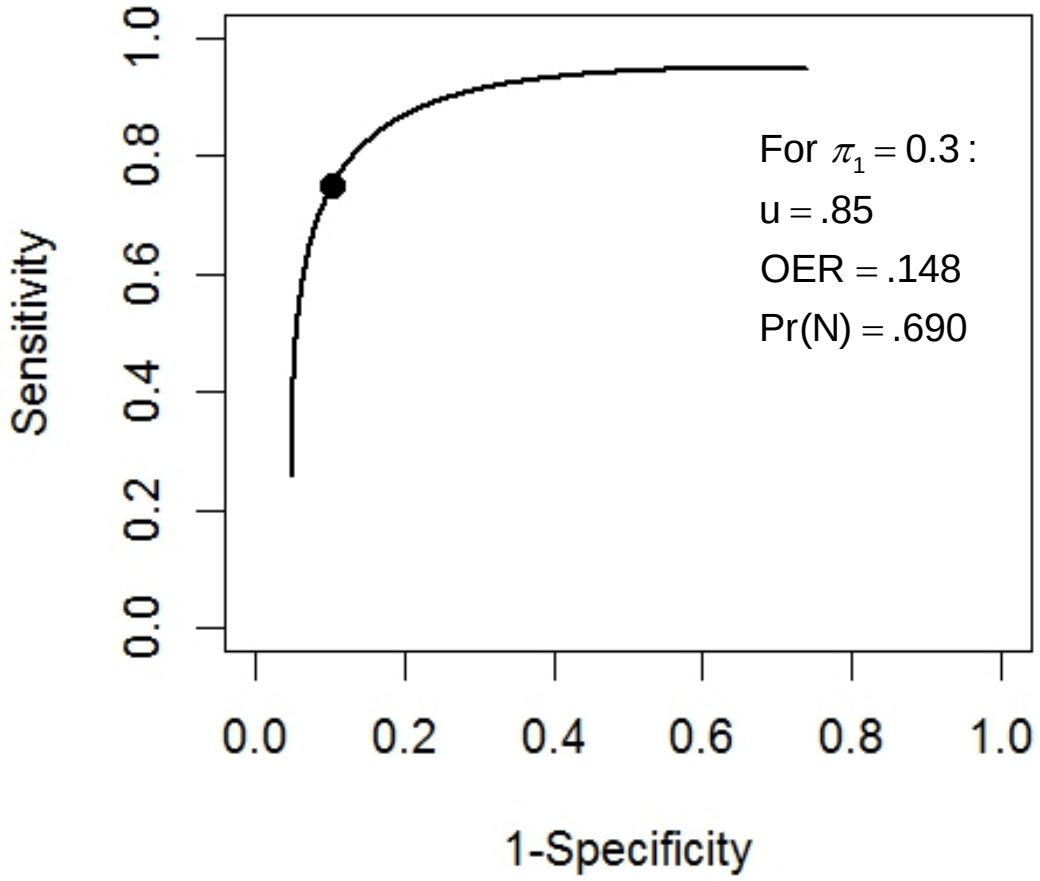


$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{C=0}{\sim} \text{MVN}\left[\begin{pmatrix} 5 \\ 2 \end{pmatrix}, \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}\right]$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{C=1}{\sim} \text{MVN}\left[\begin{pmatrix} 6 \\ 5 \end{pmatrix}, \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}\right]$$

$\alpha = \beta = .05$

Example ROC Curve of a Two-Stage Classifier



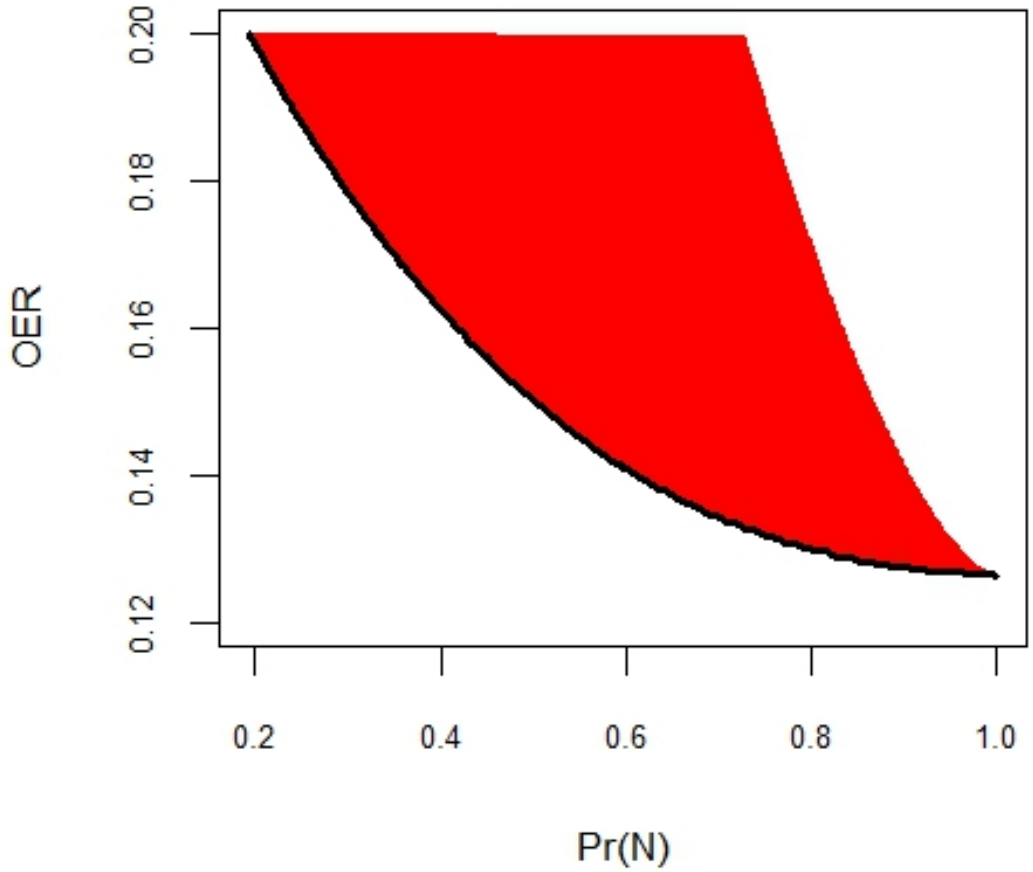
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{C=0}{\sim} \text{MVN}\left[\begin{pmatrix} 5 \\ 2 \end{pmatrix}, \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}\right]$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{C=1}{\sim} \text{MVN}\left[\begin{pmatrix} 6 \\ 5 \end{pmatrix}, \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}\right]$$

$\alpha = \beta = .05$

The optimal threshold to minimize overall error rate, for a fixed α and β, is $u = \log(\pi_0 / \pi_1)$

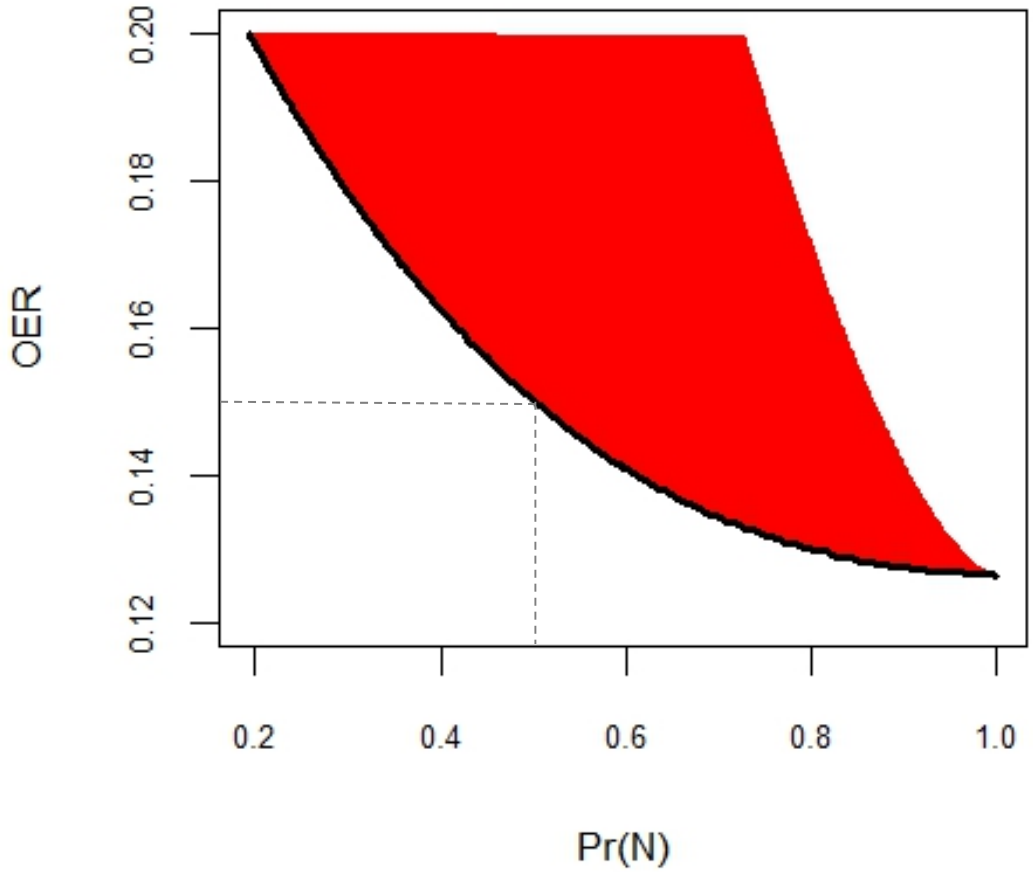
Optimizing (α,β) for a Given OER



Region of $(\text{Pr(N)}, \text{OER})$ for which $\text{OER} \leq 0.2$, obtained by varying (α, β).

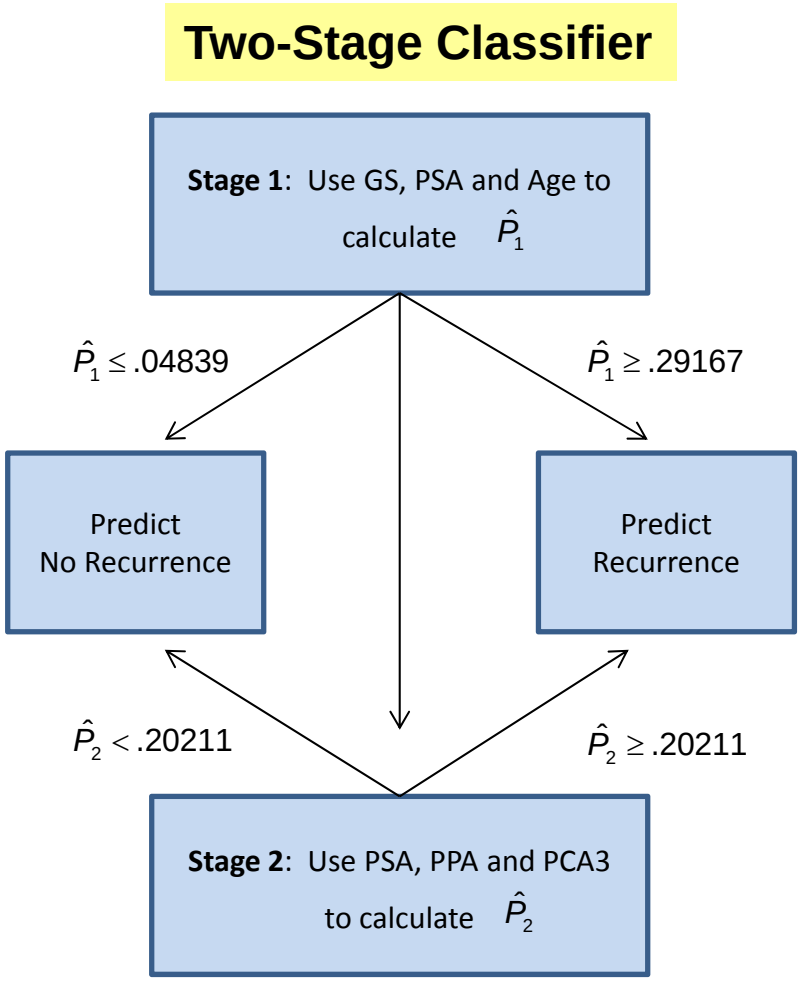
The lower boundary identifies the admissable choices for (α, β).

Optimizing (α,β) for a Given OER



For a target OER of 0.148 the admissable solution is $(\alpha\,,\beta) = (.005\,,.200)$ which delivers $\Pr(N) = .525$, which is somewhat better than the .690 obtained with the subjective choice $(\alpha\,,\beta) = (.05\,,.05)$.

Returning to the Prostate Cancer Data Set



Jeske and Smith (2017, Canadian J. of Eruology)

Quantifying the Value of the Two-Stage Classifier

Compare the two-stage classifier to the classifier that uses all of the CLIN+LAB from the beginning (comparison on training data set)

Metric	One Classifier Using All CLIN+LAB data	Two-Stage Classifier
FPR (5-fold estimate)	11.9%	9.6%
FNR (5-fold estimate)	32.1%	39.3%
Overall Accuracy (5-fold estimate)	85.6%	86.7%
NPV	95.6%	96.2%
PPV	47.1%	53.2%
Patients Referred to LAB Tests	100%	60.9%

The two classifiers are comparable with respect to the misclassification metrics, but the two-stage classifier considerably reduces the cost of testing

Summary

- Neutral zone classifiers that control FPR and FNR

Summary

- Neutral zone classifiers that control FPR and FNR
- Construction with ROC curves

Summary

- Neutral zone classifiers that control FPR and FNR
- Construction with ROC curves
- Incorporate the neutral zone into a two-stage classification framework

Summary

- Neutral zone classifiers that control FPR and FNR
- Construction with ROC curves
- Incorporate the neutral zone into a sequential classification framework
- Using the proposed method with the prostate cancer application significantly reduces the cost of treatment.

Thank You!