Classification Problems

Last modified: 2026-11-19

Source

The first section and the definition of a decision boundary moved here from Logistic Regression, whose Source note lists that page's sources. The definition of a decision region moved here from Support Vector Machines.

1 Predicting a category

Linear regression predicts a number. A great many questions are not about a number: whether a patient recovers, whether an email is spam, which of ten digits a scan shows.

Let the data be $\mathcal{X} = \{(\underset{\sim}{x}_1, y_1), \ldots, (\underset{\sim}{x}_n, y_n)\}$ with $\underset{\sim}{x}_i \in \mathbb{R}^p$ and y_i a label rather than a measurement. Two labels is binary classification; more than two is multi-class classification.

We could predict the label directly, with a classification rule[1] that maps each feature vector to one of the classes[2]. Predicting $\Pr(Y = 1 \mid \underset{\sim}{x})$ gives more information. A predicted class label names only a class; a predicted probability[3] also measures how likely each class is, so whoever uses the prediction can choose how to act on it, for instance with a threshold that reflects the cost of each kind of error.

Watch

Hutchinson teaches classification and decision boundaries[a] (6 min) [1]. The video site requires a login, which enrolled students find in the syllabus.

[a] https://facultyweb.cs.wwu.edu/~hutchib2/video_lectures/data371/#classification_overview

2 Decision regions and boundaries

A classification rule[4] splits the set[5] of all feature vectors into pieces, one for each class[6] that it predicts.

Definition 2.1 (Decision region)**.** The **decision region** of a class[a] k under a classification rule[b] c is the set[c] of feature vectors that c assigns to k:

$$\{\underset{\sim}{x} : c(\underset{\sim}{x}) = k\}.$$

[a] https://morrison-lab.github.io/sds/classification.html#def-class
[b] https://morrison-lab.github.io/sds/classification.html#def-classification-rule
[c] https://morrison-lab.github.io/mcs/sets.html#def-set

[1] https://morrison-lab.github.io/sds/classification.html#def-classification-rule
[2] https://morrison-lab.github.io/sds/classification.html#def-class
[3] https://morrison-lab.github.io/sds/classification.html#def-predicted-probability
[4] https://morrison-lab.github.io/sds/classification.html#def-classification-rule
[5] https://morrison-lab.github.io/mcs/sets.html#def-set
[6] https://morrison-lab.github.io/sds/classification.html#def-class

Example 2.1 (Decision regions of a blood-pressure rule)**.** A cohort study, the Western Collaborative Group Study [2], followed 3154 middle-aged men for coronary heart disease, and recorded each man's systolic blood pressure x, in mmHg. Let the classification rule[a] c predict heart disease ($c(x) = 1$) when $x \geq 140$, and no heart disease ($c(x) = 0$) otherwise. We chose the cutoff 140 mmHg by hand, as a round number in the upper part of the data; logistic regression instead fits a classifier[b] to data. The decision region of class[c] 1 is $\{x : x \geq 140\}$, and the decision region of class 0 is $\{x : x < 140\}$. 693 of the men are in the first region, and 2461 are in the second. Figure 1 shades the two regions over the men's blood pressures, with one bar for each recorded pressure.

```
outcome_levels <- c("No heart disease", "Heart disease")
outcome <- ifelse(wcgs$chd == 1, "Heart disease", "No heart disease")
wcgs$outcome <- factor(outcome, outcome_levels)
sbp_range <- range(wcgs$sbp)
region_names <- c("Predicted no heart disease", "Predicted heart disease")
region_starts <- c(sbp_range[1] - 1, sbp_cutoff)
region_ends <- c(sbp_cutoff, sbp_range[2] + 1)
regions <- tibble::tibble(
  xmin = region_starts,
  xmax = region_ends,
  region = region_names
)
region_fill <- c(
  "Predicted no heart disease" = "#56B4E9",
  "Predicted heart disease" = "#E69F00"
)
region_aes <- ggplot2::aes(
  xmin = xmin, xmax = xmax, ymin = -Inf, ymax = Inf, fill = region
)
panels <- ggplot2::vars(outcome)
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = sbp) +
  ggplot2::geom_rect(
    region_aes,
    data = regions,
    inherit.aes = FALSE,
    alpha = 0.3
  ) +
  ggplot2::geom_bar() +
  ggplot2::geom_vline(xintercept = sbp_cutoff, linetype = "dashed") +
  ggplot2::facet_wrap(panels, ncol = 1, scales = "free_y") +
  ggplot2::scale_fill_manual(values = region_fill) +
  ggplot2::labs(
    x = "Systolic blood pressure (mmHg)",
    y = "Number of men",
    fill = "Decision region"
  )
```

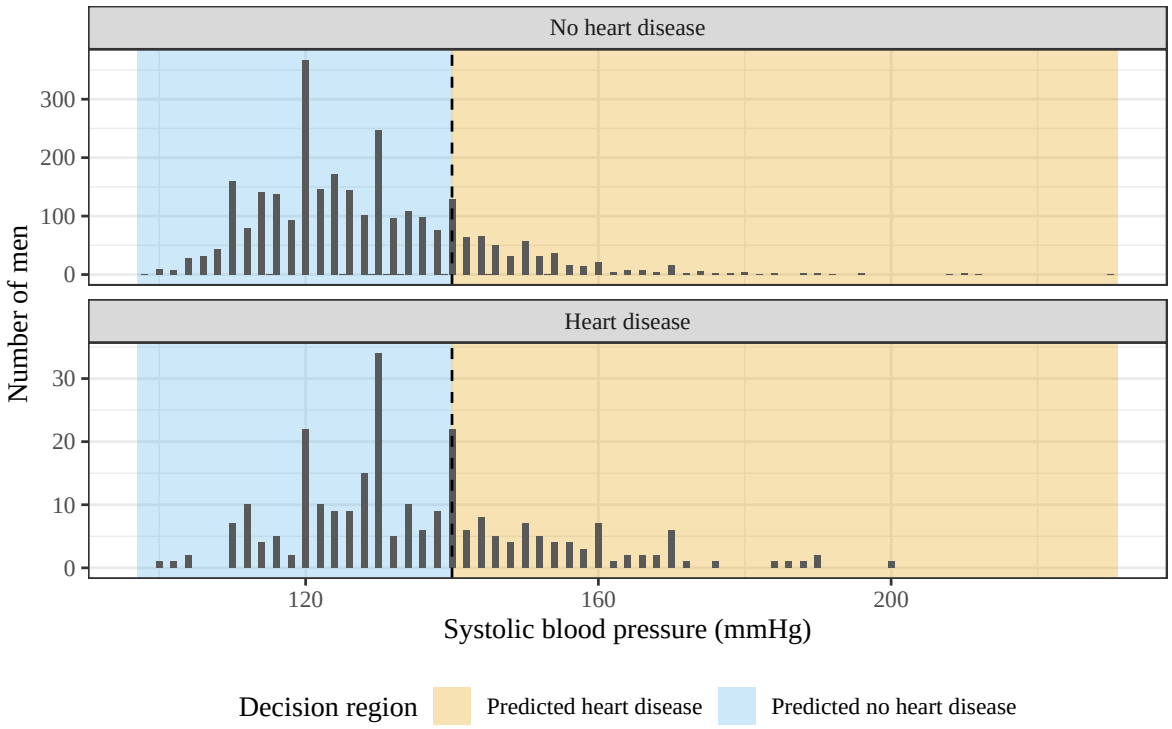


Figure 1: Number of men in the Western Collaborative Group Study at each recorded systolic blood pressure, for men without and with coronary heart disease. The shading marks the two decision regions of the rule c, and the dashed line marks its decision boundary at 140 mmHg.

[a] https://morrison-lab.github.io/sds/classification.html#def-classification-rule
[b] https://morrison-lab.github.io/sds/classification.html#def-classification
[c] https://morrison-lab.github.io/sds/classification.html#def-class

Example 2.2 (The men who had heart disease are not a decision region)**.**
In the Western Collaborative Group Study data of Example 2.1, 257 men had coronary heart disease. The set[a] of their blood pressures is not the decision region of class[b] 1 under the rule c of Example 2.1, because a decision region depends on what c predicts, not on the observed labels. Only 96 of those men are in the decision region of class 1; the other 161 had blood pressures below 140 mmHg, so c predicts no heart disease for them.

[a] https://morrison-lab.github.io/mcs/sets.html#def-set
[b] https://morrison-lab.github.io/sds/classification.html#def-class

Definition 2.2 (Decision boundary)**.** The **decision boundary** of a classifier[a] is the set of feature vectors that separates the decision regions where it predicts different classes[b].

[a] https://morrison-lab.github.io/sds/classification.html#def-classification
[b] https://morrison-lab.github.io/sds/classification.html#def-class

Example 2.3 (Decision boundary of a blood-pressure rule)**.**
The rule c of Example 2.1 predicts heart disease when the systolic blood pressure x is at least 140 mmHg. Its decision boundary is the single point $x = 140$, which lies between the decision region $\{x : x < 140\}$ and the decision region $\{x : x \geq 140\}$. It is the dashed line in Figure 1.

Example 2.4 (A cutoff that is not a decision boundary)**.**
The point $x = 120$ is not on the decision boundary of the rule c of Example 2.1. Blood pressures slightly below and slightly above 120 mmHg are both in the decision region of class[a] 0, so the point does not separate two decision regions.

[a] https://morrison-lab.github.io/sds/classification.html#def-class

3 How often a rule is wrong

The error rate[7] of a classification rule[8] counts how often it misclassifies. Example 3.1 and Example 3.2 compute it for two rules on the men of the Western Collaborative Group Study.

Example 3.1 (Error rate of a blood-pressure rule)**.**
The rule c of Example 2.1 makes two kinds of mistake on the 3154 men. It predicts heart disease for 597 men who had none, and predicts no heart disease for 161 men who had it. Its error rate[a] on these men is therefore

$$\frac{758}{3154} \approx 0.240.$$

[a] https://morrison-lab.github.io/sds/model-validation.html#def-error-rate

Example 3.2 (A rule that never predicts heart disease)**.**
A second rule c_0 predicts no heart disease for every man, so $c_0(x) = 0$ whatever x is. It is wrong only for the 257 men who had heart disease, so its error rate[a] on the 3154 men is

$$\frac{257}{3154} \approx 0.081.$$

This error rate is lower than that of the rule c in Example 3.1, even though c_0 never finds a single case of heart disease. Heart disease is rare in these data, so the error rate alone favors a rule that ignores blood pressure. The Metrics page reports each kind of mistake separately.

[a] https://morrison-lab.github.io/sds/model-validation.html#def-error-rate

4 Where the course goes next

Logistic Regression fits a classifier[9] to data, instead of choosing its cutoff by hand. The Metrics page reports how often a classifier's predictions are wrong.

References

[1] B. Hutchinson, "DATA 471/571 (machine learning) and CSCI 481/581 (deep learning) video lectures." Western Washington University. Accessed: Sep. 28, 2026. [Online]. Available: https://facultyweb.cs.wwu.edu/~hutchib2/video_lectures/data371/

[2] R. H. Rosenman, R. J. Brand, C. D. Jenkins, M. Friedman, R. Straus, and M. Wurm, "Coronary heart disease in the Western Collaborative Group Study: Final follow-up experience of 8 1/2 years," *JAMA*, vol. 233, no. 8, pp. 872–877, 1975, doi: 10.1001/jama.1975.03260080034016[10].

[7] https://morrison-lab.github.io/sds/model-validation.html#def-error-rate
[8] https://morrison-lab.github.io/sds/classification.html#def-classification-rule
[9] https://morrison-lab.github.io/sds/classification.html#def-classification
[10] https://doi.org/10.1001/jama.1975.03260080034016