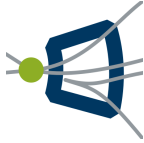


Data Analysis Block Course: Statistics and Machine learning



Exercise: Diagnosing an Infection with Temperature

Prof. Dr. Mikhail Mikhasenko

Assistants: Hendric Jonas, Christoph Tönnis



25 September 2026

[Indico schedule](#)

[RUB Moodle](#)

Learning objectives

By the end of this exercise, you should be able to

- calculate and interpret the **Gini index** for a data set,
- understand how changing a **classification threshold** affects the separation between signal and background,
- calculate **sensitivity** (true-positive rate),
- calculate **signal efficiency** and **background efficiency**,
- understand the trade-off between signal efficiency and background rejection.

Scenario

A hospital is testing whether a patient's **body temperature** can be used to diagnose an infection.

Ten patients have been examined. After a more definitive medical test, we know whether each patient actually had an infection. We will call

- **Signal:** patients who actually have an infection.
- **Background:** patients who do not have an infection.

The hospital considers a patient **positive for infection** if their temperature is **above a chosen threshold**.

Patient data

Patient	Temperature (°C)	Actually infected?
1	36.5	No
2	36.8	No
3	37.0	No
4	37.2	Yes
5	37.5	No
6	37.7	Yes
7	38.0	Yes
8	38.2	Yes
9	38.5	No
10	39.0	Yes

There are therefore **5 infected patients (signal)** and **5 non-infected patients (background)**.

Exercise 1 - Choosing a temperature threshold

Suppose the hospital initially chooses the selection

$$T > 37.5^{\circ}\text{C}$$

A patient is classified as **positive** when their temperature is above this value.

1. Which patients are classified as positive?
2. Among these patients, how many are
 - **True positives (TP)** - infected and classified as positive?
 - **False positives (FP)** - not infected but classified as positive?
3. How many patients are classified as negative?
4. Among the negative patients, how many are
 - **True negatives (TN)**?
 - **False negatives (FN)**?

Complete the confusion matrix:

	Actually infected	Not infected
Classified positive	TP = ____	FP = ____
Classified negative	FN = ____	TN = ____

Exercise 2 - Sensitivity

Sensitivity measures the fraction of all truly infected patients that the diagnostic method successfully identifies. It is defined as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

Using the threshold $T > 37.5^{\circ}\text{C}$:

1. Calculate the sensitivity.
2. Express your answer as a percentage.
3. What does this percentage mean in the context of the hospital?

Exercise 3 - Signal and background efficiency

In a classification problem, we can interpret the infected patients as **signal** and the non-infected patients as **background**.

Signal efficiency

The **signal efficiency** is the fraction of all signal events that pass the selection

$$\epsilon_S = \frac{\text{number of signal events passing}}{\text{total number of signal events}} \quad (2)$$

In this example we have

$$\epsilon_S = \text{sensitivity}$$

Background efficiency

The **background efficiency** is the fraction of all background events that pass the selection

$$\epsilon_B = \frac{\text{number of background events passing}}{\text{total number of background events}} \quad (3)$$

For the threshold $T > 37.5^\circ\text{C}$:

1. Calculate the signal efficiency.
2. Calculate the background efficiency.
3. What fraction of the background is rejected?
4. Is it possible to increase the signal efficiency while decreasing the background efficiency? Investigate this by changing the temperature threshold.

Exercise 4 - The Gini index

The **Gini index** is a measure of how mixed two classes are within a sample. For two classes, it is defined as

$$G = 1 - p_S^2 - p_B^2 \quad (4)$$

where p_S is the fraction of patients who are signal, p_B is the fraction of patients who are background, and $p_S + p_B = 1$.

A sample containing only one class has $G = 0$, while a sample containing equal amounts of signal and background has $G = 0.5$.

Consider **all 10 patients together**.

1. What is p_S ?
2. What is p_B ?
3. Calculate the Gini index. Now consider the patients classified as **positive** using $T > 37.5^\circ\text{C}$.

4. How many signal patients are in this group?
5. How many background patients are in this group?
6. Calculate p_S and p_B for this group.
7. Calculate its Gini index. Finally, consider the patients classified as **negative**.
8. Calculate the Gini index for the negative group.
9. Compare the Gini indices of the positive and negative groups with the Gini index of the original sample.
10. What does a lower Gini index tell you about the composition of a group?

Exercise 5 - Investigating the threshold

The choice of 37.5°C was arbitrary. Let's see what happens when we change it. Calculate the signal and background efficiencies for several thresholds.

1. What happens to the **signal efficiency** as the threshold is increased?
2. What happens to the **background efficiency**?
3. Why is there a trade-off between keeping signal and rejecting background?
4. Which threshold gives the largest signal efficiency?
5. Which threshold gives the smallest background efficiency?
6. Is there a single threshold that simultaneously maximizes signal efficiency and minimizes background efficiency? Explain.

Exercise 6 - Finding a useful separation

Challenge: Suppose the hospital wants a diagnostic method that keeps as many infected patients as possible while rejecting as many non-infected patients as possible.

One way of visualizing the performance is to plot signal efficiency against background efficiency. Using the results from Exercise 5:

1. Make a scatter plot with **background efficiency on the x-axis** and **signal efficiency on the y-axis**.
2. Put a label next to each point showing its temperature threshold.
3. Describe what happens as the temperature threshold is increased.
4. Explain why the choice of threshold depends on the consequences of false positives versus false negatives.

Short conceptual questions

1. If missing an infected patient is considered very serious, would the hospital tend to prefer a **higher or lower** temperature threshold?
2. What would happen to the number of false positives if the threshold were lowered?
3. What would happen to the sensitivity?
4. Why might a hospital not simply choose the threshold that gives the smallest background efficiency?
5. How does this example illustrate the general problem of **classification** in machine learning and experimental physics?

Key concepts

Concept	Meaning in this exercise
Signal	Patients who actually have an infection
Background	Patients who do not have an infection
Threshold	Temperature used to classify a patient as positive
TP	Infected patient correctly classified as positive
FP	Non-infected patient incorrectly classified as positive
FN	Infected patient incorrectly classified as negative
TN	Non-infected patient correctly classified as negative
Sensitivity	Fraction of infected patients identified
Signal efficiency	Fraction of signal passing the selection
Background efficiency	Fraction of background passing the selection
Gini index	Measure of how mixed the two classes are