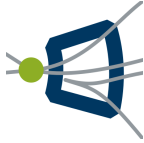


Data Analysis Block Course: Statistics and Machine learning



Exercise: Diagnosing an Infection with Temperature - Solutions

Prof. Dr. Mikhail Mikhasenko

Assistants: Hendric Jonas, Christoph Tönnis



25 September 2026

[Indico schedule](#)

[RUB Moodle](#)

Solution 1 - Choosing a temperature threshold

The selection is $T > 37.5^\circ\text{C}$. Because the inequality is strict, patient 5 at exactly 37.5°C is classified as negative.

1. The positive patients are **6, 7, 8, 9 and 10**.
2. Patients 6, 7, 8 and 10 are infected, so there are **4 true positives**. Patient 9 is not infected, giving **1 false positive**.
3. The negative patients are **1, 2, 3, 4 and 5**, so there are **5 negative classifications**.
4. Patients 1, 2, 3 and 5 are not infected, giving **4 true negatives**. Patient 4 is infected, giving **1 false negative**.

The completed confusion matrix is

	Actually infected	Not infected
Classified positive	TP = 4	FP = 1
Classified negative	FN = 1	TN = 4

Solution 2 - Sensitivity

Sensitivity is defined as

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

For the threshold $T > 37.5^\circ\text{C}$, this gives

$$\text{Sensitivity} = \frac{4}{4 + 1} = \frac{4}{5} = 0.80 = 80\%$$

Thus **80% of the truly infected patients are correctly identified**. Four of the five infected patients test positive, while one is missed. This percentage refers to all infected patients, not to all ten patients or to all positive classifications.

Solution 3 - Signal and background efficiency

Using the same selection, the efficiencies are

$$\begin{aligned}\epsilon_S &= \frac{4}{5} = 80\% \\ \epsilon_B &= \frac{1}{5} = 20\% \\ 1 - \epsilon_B &= 80\%\end{aligned}$$

The last line is the **background rejection**: four of the five background patients are rejected. In this example, signal efficiency is the same as sensitivity.

Can both efficiencies improve at once?

With a single selection $T > t$, lowering t can only add patients to the selected group; raising t can only remove patients. Consequently, both efficiencies are non-increasing functions of t . It is **not possible to strictly increase signal efficiency while strictly decreasing background efficiency** by changing only this threshold.

For example, lowering the threshold from 37.5°C to 37.0°C increases signal efficiency from 80% to 100%, but also increases background efficiency from 20% to 40%. Raising it to 38.5°C reduces background efficiency to 0%, but signal efficiency falls to 20%.

One efficiency can remain unchanged while the other changes. For example, raising the threshold from 36.5°C to 37.0°C removes only background patients: signal efficiency stays at 100%, while background efficiency falls from 80% to 40%.

Solution 4 - The Gini index

For two classes, the Gini index is

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

All 10 patients

There are five signal and five background patients, so

$$\begin{aligned}p_S &= \frac{5}{10} = 0.5 \\ p_B &= \frac{5}{10} = 0.5 \\ G &= 1 - (0.5)^2 - (0.5)^2 = 0.5\end{aligned}$$

Positive group

At $T > 37.5^\circ\text{C}$, four signal patients and one background patient pass the selection. Thus

$$\begin{aligned}p_S &= \frac{4}{5} = 0.8 \\ p_B &= \frac{1}{5} = 0.2 \\ G_{\text{positive}} &= 1 - (0.8)^2 - (0.2)^2 = 0.32\end{aligned}$$

Negative group

The negative group contains one signal patient and four background patients. Thus

$$\begin{aligned}p_S &= \frac{1}{5} = 0.2 \\p_B &= \frac{4}{5} = 0.8 \\G_{\text{negative}} &= 1 - (0.2)^2 - (0.8)^2 = 0.32\end{aligned}$$

Both groups have the same Gini index, 0.32, which is lower than the original sample's value of 0.5. A **lower Gini index means a purer group**: one class is more dominant. Here the positive group is mostly signal and the negative group is mostly background. The Gini index does not distinguish which class dominates.

Solution 5 - Investigating the threshold

For each cut $T > t$, count the selected infected and non-infected patients and divide by the five patients in the corresponding class. All efficiencies and rejected fractions in the table are percentages.

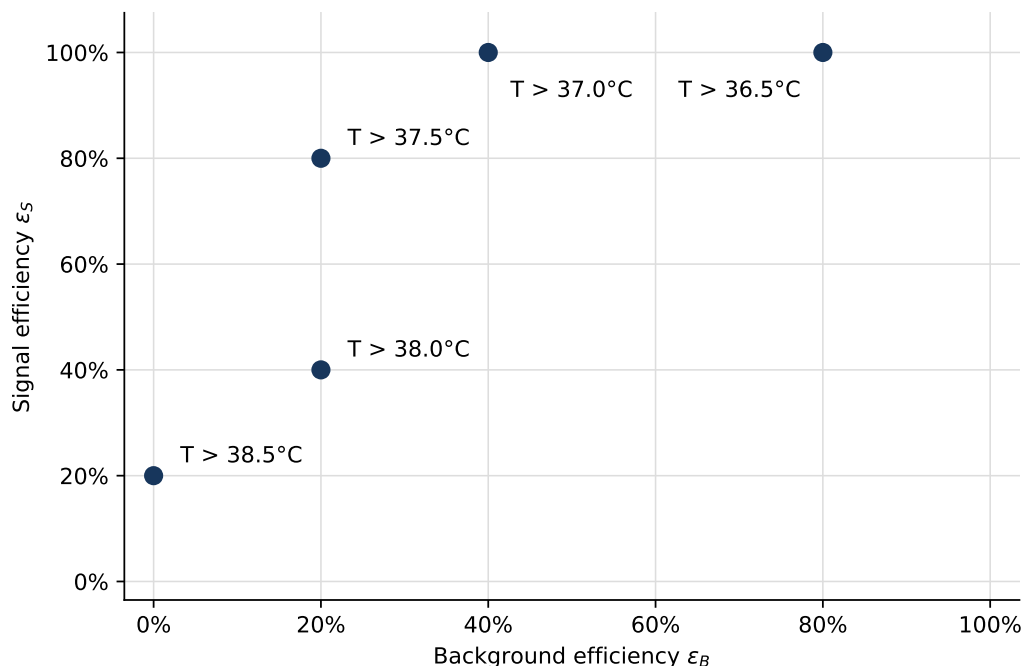
t (°C)	ϵ_S	ϵ_B	$1 - \epsilon_S$	$1 - \epsilon_B$
36.5	100%	80%	0%	20%
37.0	100%	40%	0%	60%
37.5	80%	20%	20%	80%
38.0	40%	20%	60%	80%
38.5	20%	0%	80%	100%

For $T > 37.0^\circ\text{C}$, only patients 5 and 9 are selected background patients. Patient 3 at exactly 37.0°C is excluded. The background efficiency is therefore $2/5 = 40\%$.

1. **Signal efficiency** is non-increasing as the threshold rises. It decreases when an infected patient is removed and otherwise stays constant.
2. **Background efficiency** is also non-increasing. It decreases when a non-infected patient is removed and otherwise stays constant.
3. The temperatures of the two classes overlap. Raising the threshold can reject background but can also remove signal. Keeping all signal therefore comes at the cost of accepting some background.
4. The largest signal efficiency in the table is 100%, attained at $t = 36.5^\circ\text{C}$ and $t = 37.0^\circ\text{C}$. More generally, every $t < 37.2^\circ\text{C}$ keeps all five infected patients.
5. The smallest background efficiency in the table is 0%, attained at $t = 38.5^\circ\text{C}$. More generally, every $t \geq 38.5^\circ\text{C}$ rejects all background patients.
6. **No threshold attains both optima.** Keeping all signal requires $t < 37.2^\circ\text{C}$, while rejecting all background requires $t \geq 38.5^\circ\text{C}$. These conditions cannot hold simultaneously.

Solution 6 - Finding a useful separation

Plot background efficiency on the horizontal axis and signal efficiency on the vertical axis. Each point below is labelled with its selection threshold from Solution 5.



As the threshold increases from 36.5°C to 38.5°C , the selected points move leftwards and downwards, with horizontal or vertical steps when only one class is affected. The top-left corner would represent perfect separation: $\epsilon_S = 1$ and $\epsilon_B = 0$. None of these thresholds reaches it, and the incompatible conditions in Solution 5 show that no other temperature threshold can do so for this data set.

The choice depends on the consequences of the two errors. If missing an infected patient is particularly costly, a lower threshold retains more signal. If unnecessary follow-up tests or treatment after a false positive are particularly costly, a higher threshold may be preferred. The data alone do not specify a unique best threshold without these priorities.

Solutions - Short conceptual questions

1. The hospital would tend to prefer a **lower threshold**, because this reduces the number of missed infected patients and increases or preserves sensitivity.
2. Lowering the threshold can only increase or preserve the number of **false positives**. It cannot reduce that number.
3. Sensitivity increases or stays constant when the threshold is lowered. It increases whenever an additional infected patient passes the selection.
4. Minimizing background efficiency alone can miss many infected patients. For example, $T > 38.5^\circ\text{C}$ gives zero background efficiency but identifies only one of the five infected patients. For $T > 39.0^\circ\text{C}$, nobody is selected at all.
5. This illustrates **binary classification**: a measured quantity and a decision threshold separate two overlapping classes. Changing the threshold changes true-positive and false-positive rates, and the preferred balance depends on the consequences of the errors. The same reasoning applies to signal and background in experimental physics.