Lecture: Logistic Regression

Binary classification, sigmoid and model tuning

K. Kadri

Outline

Problem: predicting a binary class

- We want to predict a variable $y \in \{0, 1\}$ (e.g., *sick* / *not sick*).
- A linear regression would output an unbounded real value.
- We need a model that outputs a **probability** between 0 and 1:

$$p(y = 1 \mid \mathrm{x}) \in [0, 1]$$

Key idea

Logistic regression does not directly predict a class, but the **probability of the positive class**, then applies a **threshold** (often 0.5).

The sigmoid function

Definition

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where $z = \mathrm{w}^\top \mathrm{x} + b$.

- For $z \to +\infty$: $\sigma(z) \to 1$.
- For $z \to -\infty$: $\sigma(z) \to 0$.
- Interpretation: $\sigma(z)$ is a **probability**.

Logistic model

$$p(y = 1 \mid \mathrm{x}) = \sigma(\mathrm{w}^\top \mathrm{x} + b)$$

Log-odds and linear model

- The model can be rewritten as **log-odds**:

$$\log \frac{p}{1-p} = \mathrm{w}^\top \mathrm{x} + b$$

- $\dfrac{p}{1-p}$: **odds** ratio.
- The algorithm learns w and b that explain this odds ratio.

Interpretation

Each coefficient w_j measures the effect of feature x_j on the **log-odds** of belonging to the positive class.

Cost function: log-loss

For one example (x_i, y_i)

$$\mathcal{L}(y_i, \hat{p}_i) = -\Big[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\Big]$$

where $\hat{p}_i = p(y = 1 \mid \mathrm{x}_i)$.

- If the model is confident and correct $\rightarrow$ low loss.
- If the model is confident and wrong $\rightarrow$ very high loss.

Learning objective

Minimize the sum (or mean) of the log-loss over the entire training set.

L2 and L1 regularization

- **Problem**: overfitting if coefficients have too much freedom.
- **L2 regularization** (Ridge):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{log-loss}} + \lambda \|w\|_2^2$$

- **L1 regularization** (Lasso):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{log-loss}} + \lambda \|w\|_1$$

Effect

- L2: shrinks all coefficients (rarely to 0).
- L1: pushes some coefficients exactly to 0 $\rightarrow$ **feature selection**.

Hyperparameter C in scikit-learn

In `LogisticRegression`

$$C = \frac{1}{\lambda}$$

- Small $C \rightarrow$ strong regularization (simpler, more biased model).
- Large $C \rightarrow$ weak regularization (more complex, higher overfitting risk).

Tuning

We choose C (and penalty type) via **cross-validation** (GridSearch, RandomizedSearch).

Decision threshold

- Default:

$$\hat{y} = \begin{cases} 1 & \text{if } p(y = 1 \mid \mathrm{x}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

- But we can **adapt the threshold** depending on the problem:
  - lower threshold $\rightarrow$ more positives detected (recall $\uparrow$, FP risk $\uparrow$).
  - higher threshold $\rightarrow$ fewer false positives but lower recall.

Link with metrics

We choose the threshold based on the **trade-off** Precision / Recall / F1-score and the **business cost** of false positives and false negatives.

ROC curve and AUC

- By varying the threshold, we obtain multiple (FPR, TPR) points.
- The ROC curve plots:

$$\text{TPR} \quad \text{as a function of} \quad \text{FPR}$$

- The area under the curve (AUC) measures the model's **overall ability** to separate classes.

Rule of thumb

- AUC ≈ 0.5: random model.
- AUC close to 1: very good separation.

Advantages of logistic regression

- **Simple** model, fast to train.
- **Probabilistic** output, easy to interpret.
- Interpretable: coefficients can be analyzed.
- Works well as a **baseline** on many problems.

Typical use cases

- Credit scoring.
- Customer churn probability.
- Binary diagnosis (presence/absence of a disease).

Limitations and best practices

- Assumes a **linear decision boundary** in feature space.
- Sensitive to **poorly scaled** features $\rightarrow$ standardization important.
- May be insufficient if the relationship is highly nonlinear.

Best practices

- Always standardize continuous features.
- Test several C values and penalty types.
- Examine coefficients and metrics (F1, AUC) rather than accuracy alone.