Mar 20, 2023

Machine Learning Methods and Applications

Week 4. Model validation and pre-processing

Remember

- c-dependent performance metrics
- do not use only one metric to evaluate model performance
- categorical variables with many classes
- missing values
- imbalanced classes of feature

Validation methods

Model validation

- Model validation ~ Model performance check

- Data splitting (train and test) provides only one estimate of model performance.

- Resampling based validation techniques for exploring model performance, provides k estimates of model performance during the model training phase.

Cross-validation methods

1. Holdout cross-validation
2. K-fold cross-validation
3. Stratified cross-validation
4. Leave-p-out cross-validation
5. Leave-one-out cross-validation

1. Holdout cross-validation

- The classical splitting the data as train and test data.

2. K-fold cross-validation

- Training data is randomly splitted into k sets of roughly equal size.

- Folds are used to perform k iterations of model fitting and evaluation.

- 5-fold CV is the five estimates of model performance in total.

2. K-fold cross-validation

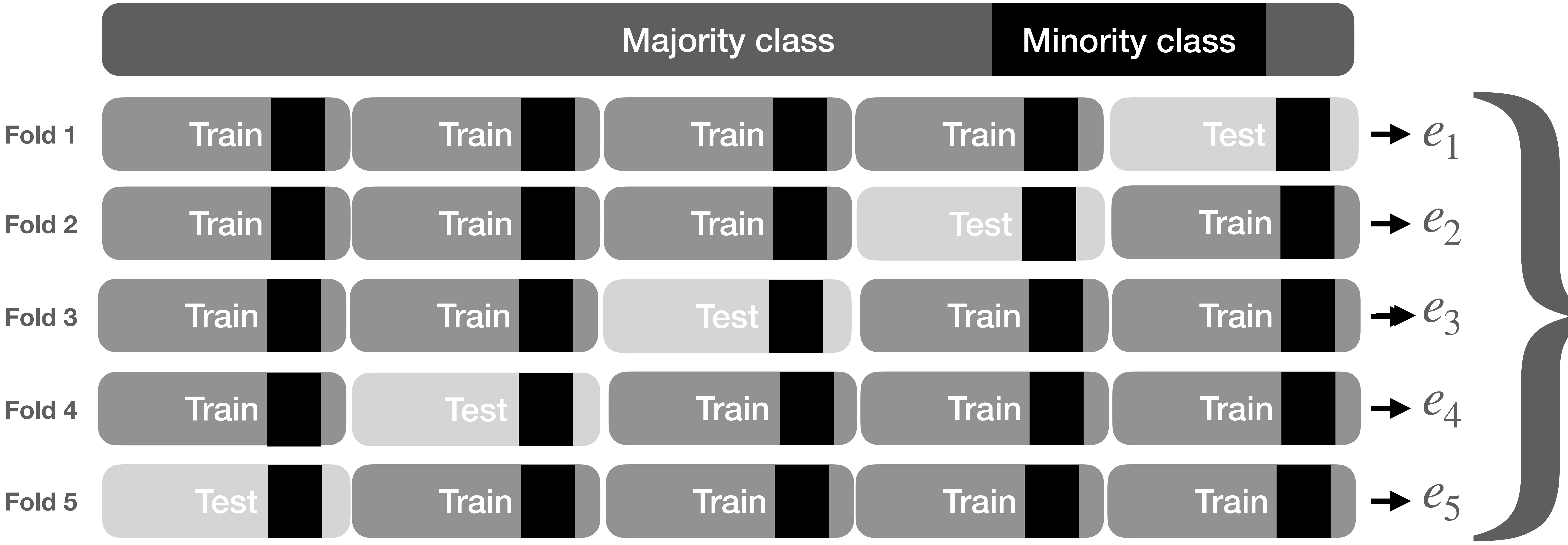
	All data					
Fold 1	Train	Train	Train	Train	Test	$\rightarrow e_1$
Fold 2	Train	Train	Train	Test	Train	$\rightarrow e_2$
Fold 3	Train	Train	Test	Train	Train	$\rightarrow e_3$
Fold 4	Train	Test	Train	Train	Train	$\rightarrow e_4$
Fold 5	Test	Train	Train	Train	Train	$\rightarrow e_5$

$$e_t = \frac{1}{5} \sum_{i=1}^{k} e_i$$

3. Stratified cross-validation

- Useful for class imbalance problem in target variable
- Similar to the k-fold cross-validation
- The difference is to consider the ratio of majority and minority classes in splitting.

3. Stratified cross-validation



$$e_t = \frac{1}{5}\sum_{i=1}^{k} e_i$$

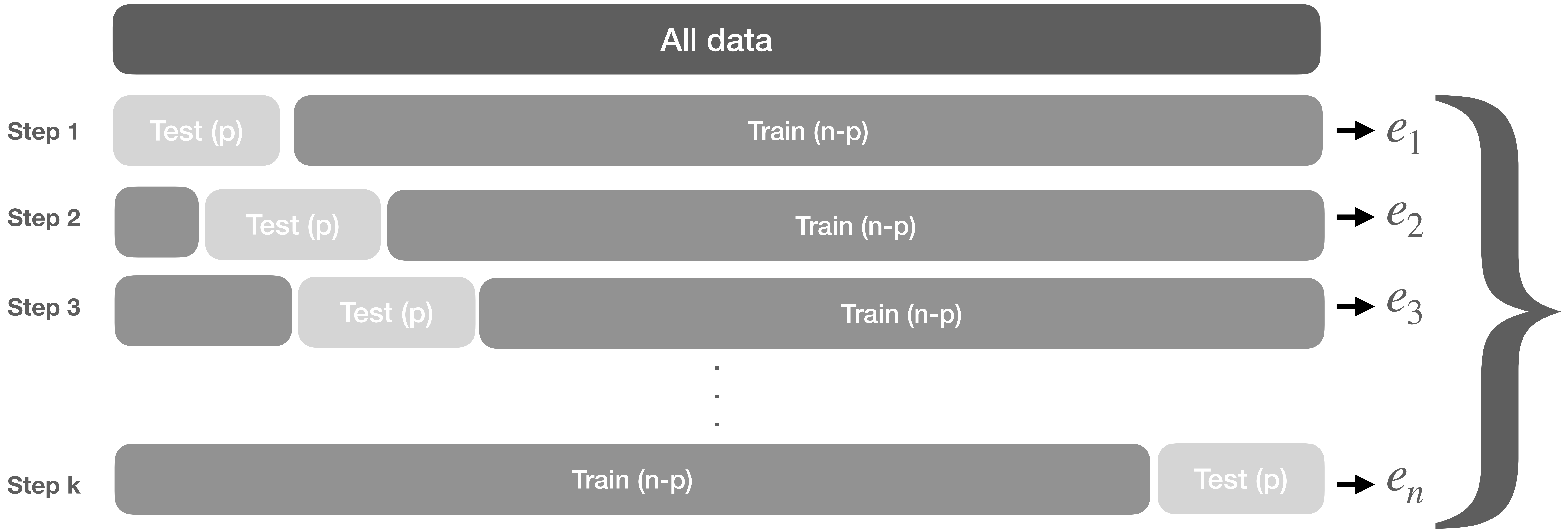
4. Leave-p-out cross-validation (LpOC)

K-fold cross-validation overlaps for LpOC if $p > 1$.

Steps:

1. Choose p samples from the all data as test data
2. The remaining n-p samples will be train data
3. Repeat the two steps above x times
4. To get the final performance average the results on each step

4. Leave-p-out cross-validation



$$e_t = \frac{1}{k}\sum_{i=1}^{k} e_i$$

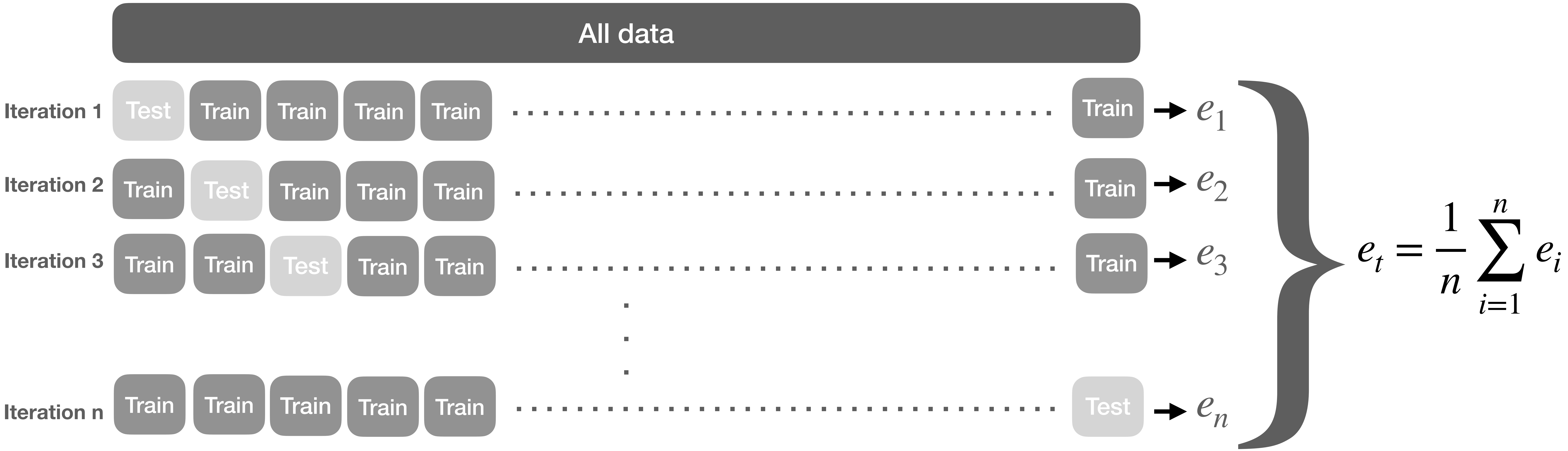
5. Leave-one-out cross-validation (LOOCV)

It is extreme case of k-fold cross-validation ($k = n$) and the special case of leave-p-out cross-validation ($p = 1$).

Steps:

1. Choose one sample from the all data which will be the test data
2. The remaining n-1 samples will be train data
3. Train the model on train data
4. Validate the model performance on test data
5. Save the results
6. Repeat the steps above for n (sample size) times
7. To get the final score average the results.

5. Leave-one-out cross-validation



Missing data problem

Missing data

- Missing (or NA: non-available) values in observations may be problematic in predictive models.
- Even some of the models are not trained in case of missing data.

Missing data

Handling missing data:

1. **Remove**, discard observations with missing values from the data set.
2. **Impute**, “fill in” the missing values with other values.

Missingness mechanisms

1. Missing completely at random (MCAR)
2. Missing at random (MAR)
3. Missing not at random (MNAR)

1. Missing completely at random (MCAR)

Missingness has no association with any data you have observed or not.

- Imputation is suggested.
- Removing observations may reduce sample size (or loss of information), but will not bias.

2. Missing at random (MAR)

Missingness depends on observed data, but not the unobserved data.

- Imputation is suggested.
- Removing observations not ideal, may lead to bias.

3. Missing not at random (MNAR)

Missingness is related to an unobserved value relevant to the assessment of interest.

- Data will be biased from removing and imputation.
- Inference can be limited, proceed with caution.

Transformations

Transformation

- Data transformation in ML is also called **feature scaling**
- It is used to scale the features in different scales (a.k.a. ranges)
- **Not always needed**
- Performance of some models may be improved after scaling especially in unsupervised learning

Min-max scaling

- It maps a numerical value x to the [0, 1] interval

$$x_t = \frac{x - min(x)}{max(x) - min(x)}$$

Normalization

- Also called **standardization**
- It maps a numerical value x to a new distribution with $\mu = 0$ and standard deviation $\sigma = 1$

$$x_t = \frac{x - mean(x)}{sd(x)}$$

Min-max scaling vs. Normalization

Min-max scaling

> Ensures that all features share the exact same scale

> Does not handle well with outliers

Normalization

> More robust to outliers

> Normalized data may be on different scales

Application

See the R codes on the course GitHub repository!

The video recording of today's lecture will be available on **YouTube**, and slides on **GitHub**.
Feel free to contact me via e-mail: mustafacavus@eskisehir.edu.tr