

Machine learning Part 2

Model selection & validation

Jérôme Dockès & Nikhil Bhagwat

MAIN tutorials course 2021-11-26

Outline

Introduction: cross-validation

Model and hyperparameter selection

Dimensionality reduction

Conclusion

Recap of part 1

Supervised learning

- Regression: least-squares linear regression
- Classification: logistic regression

Linear regression



Logistic regression



Recap of part 1

Supervised learning

- Regression: least-squares linear regression
- Classification: logistic regression

Regularization

- ℓ_2 a.k.a. ridge regularization

Recap of part 1

Supervised learning

- Regression: least-squares linear regression
- Classification: logistic regression

Regularization

- ℓ_2 a.k.a. ridge regularization

Model evaluation and selection

- Out-of-sample generalization; independent test set
- Performance metrics:
 - regression: mean squared error
 - classification: accuracy, ROC curve
- Cross-validation

Notation & vocabulary

Supervised learning framework

$$Y = f(X) + E \quad (1)$$

- $Y \in \mathbb{R}$: output (a.k.a. target, dependent variable) to predict

Notation & vocabulary

Supervised learning framework

$$Y = f(X) + E \quad (1)$$

- $Y \in \mathbb{R}$: output (a.k.a. target, dependent variable) to predict
- $X \in \mathbb{R}^p$: features (a.k.a. inputs, regressors, descriptors, independent variables)

Notation & vocabulary

Supervised learning framework

$$Y = f(X) + E \quad (1)$$

- $Y \in \mathbb{R}$: output (a.k.a. target, dependent variable) to predict
- $X \in \mathbb{R}^p$: features (a.k.a. inputs, regressors, descriptors, independent variables)
- $E \in \mathbb{R}$: unmodelled noise

Notation & vocabulary

Supervised learning framework

$$Y = f(X) + E \quad (1)$$

- $Y \in \mathbb{R}$: output (a.k.a. target, dependent variable) to predict
- $X \in \mathbb{R}^p$: features (a.k.a. inputs, regressors, descriptors, independent variables)
- $E \in \mathbb{R}$: unmodelled noise
- f : the function we try to approximate

Example: linear regression

$$Y = \beta_0 + \langle X, \beta \rangle + E \quad (2)$$

$$= \beta_0 + \sum_{j=1}^p X_j \beta_j + E \quad (3)$$

"learning" = estimating $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$

How to estimate parameters: Empirical Risk Minimization

Given n training examples $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$,
find $\hat{\beta}_0 \in \mathbb{R}$, $\hat{\beta} \in \mathbb{R}^p$
that minimize the empirical risk:

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \|\mathbf{y} - \hat{\beta}_0 - \mathbf{X} \hat{\beta}\|_2^2 \quad (4)$$

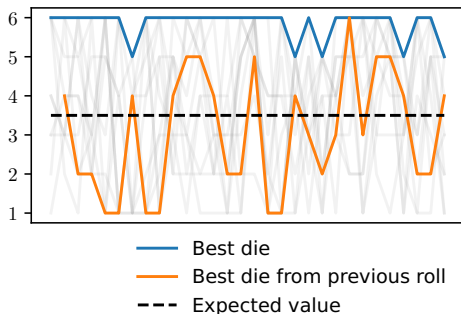
$$= \sum_{i=1}^n (\mathbf{y}_i - \hat{\beta}_0 - \sum_{j=1}^p \mathbf{x}_{ij} \hat{\beta}_j)^2 \quad (5)$$

"Fitting" the parameters to \mathbf{X}, \mathbf{y} .

Need for fresh test data

When you hear "best", "maximum", "select", ... think "bias"

- I have 4 dice and want to find one that rolls high numbers
- I roll them all once and select the die that gives the highest number
- The selected die rolled a 5. Is 5 a good estimate of that die's average result? What if I had 1,000 dice?
- I need to roll it again to get an unbiased estimate



Estimating prediction performance

When you hear "best", "maximum", "select", ... think "bias"

Setting the parameters

- **Select** β that gives the **best** prediction on training data
- The prediction score for $\hat{\beta}$ is biased: compute a new score on unseen test data.

scikit-learn "estimator API": fit; predict

```
estimator = Ridge()  
estimator.fit(X_train, y_train)  
predictions = estimator.predict(X_test)
```

https://scikit-learn.org/stable/getting_started.html
`sklearn.linear_model.Ridge`











Evaluating performance with `sklearn.metrics`

```
estimator = Ridge()  
estimator.fit(X_train, y_train)  
predictions = estimator.predict(X_test)  
  
mse = metrics.mean_squared_error(y_test, predictions)
```

https://scikit-learn.org/stable/getting_started.html
`sklearn.linear_model.Ridge`
`sklearn.metrics`
more info on model evaluation

`ex_01_fit_predict.py`

Cross-validation

Fold 0	Train		Score 0
	Test		
Fold 1	Train		Score 1
	Test		
Fold 2	Train		Score 2
	Test		
Fold 3	Train		Score 3
	Test		
Fold 4	Train		Score 4
	Test		

scikit-learn.org/stable/modules/cross_validation.html

`sklearn.model_selection.cross_validate`

`ex_02_cross_validate.py`

Outline

Introduction: cross-validation

Model and hyperparameter selection

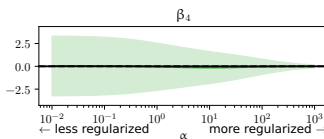
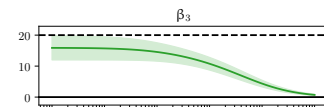
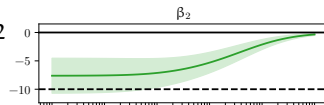
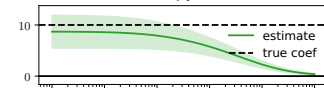
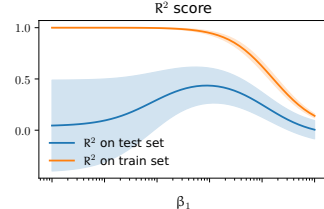
Dimensionality reduction

Conclusion

Regularization

Example: ridge regression

$$\operatorname{argmin}_{\beta, \beta_0} \|\mathbf{y} - \beta_0 - \mathbf{X} \beta\| + \alpha \|\beta\|_2^2 \quad (6)$$



$$\text{Var}(\hat{\beta}_i) = \mathbb{E}(\hat{\beta}_i - \mathbb{E}(\hat{\beta}_i))^2$$

$$\text{Bias}(\hat{\beta}_i) = \mathbb{E}(\hat{\beta}_i) - \beta_i$$

Setting hyperparameters

How can we choose the ridge hyperparameter α ?

Try a few and pick the best one...

But measure its performance on separate data!

Nested cross-validation

When you hear "best", "maximum", "select", ... think "bias"

Nested cross-validation

When you hear "best", "maximum", "select", ... think "bias"

Setting the parameters

- **Select** β that gives the **best** prediction on training data
- The prediction score for $\hat{\beta}$ is biased: compute a new score on unseen test data.

Nested cross-validation

When you hear "best", "maximum", "select", ... think "bias"

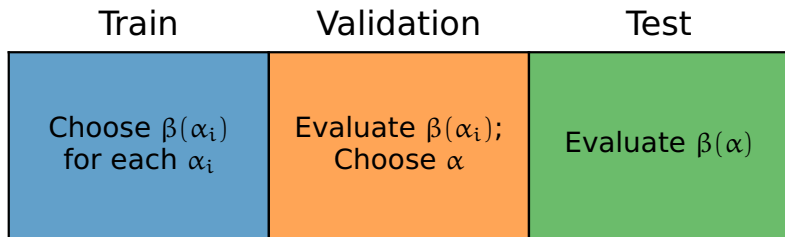
Setting the parameters

- **Select** β that gives the **best** prediction on training data
- The prediction score for $\hat{\beta}$ is biased: compute a new score on unseen test data.

Setting the hyperparameters

- Repeat step 1 for a few values of α , fitting and testing several models
- **Select** the hyperparameter that obtains the **best** prediction on test data
- The prediction score of that model on *test* data is biased: evaluate it again on unseen data

One split



Nested cross-validation

Fold 0	Train	Fold 0	Train	For all α	
			Test	For all α	
		Fold 1	Train	For all α	
			Test	For all α	
		Fold 2	Train	For all α	
			Test	For all α	
		Refit		For best α	
	Test				

Score 0

Fold 1	Train	Fold 0	Train	For all α	
			Test	For all α	
		Fold 1	Train	For all α	
			Test	For all α	
		Fold 2	Train	For all α	
			Test	For all α	
		Refit		For best α	
	Test				

Score 1

Fold 2	Train	Fold 0	Train	For all α	
			Test	For all α	
		Fold 1	Train	For all α	
			Test	For all α	
		Fold 2	Train	For all α	
			Test	For all α	
		Refit		For best α	
	Test				

Score 2

Fold 3	Train	Fold 0	Train	For all α	
			Test	For all α	
		Fold 1	Train	For all α	
			Test	For all α	
		Fold 2	Train	Model and hyperparameter	
			Test	Model and hyperparameter	

Implementing nested CV

`ex_05_nested_cross_validation.py`

Outline

Introduction: cross-validation

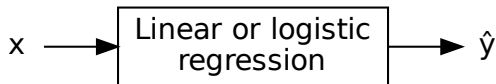
Model and hyperparameter selection

Dimensionality reduction

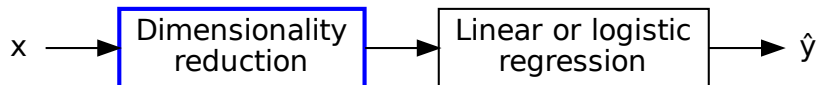
Conclusion

Dimensionality reduction

Until now



Add a step in the pipeline: simplifying the inputs



Dimensionality reduction

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{X}_{:,1} + \hat{\beta}_2 \mathbf{X}_{:,2} + \cdots + \hat{\beta}_p \mathbf{X}_{:,p} \quad (7)$$

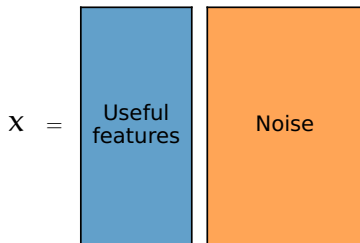
Problems when the number of features p becomes large

- Bigger errors on test data (larger variance of predictions)
- Numerical stability issues
- Computational cost and memory usage

Simulated data for linear regression

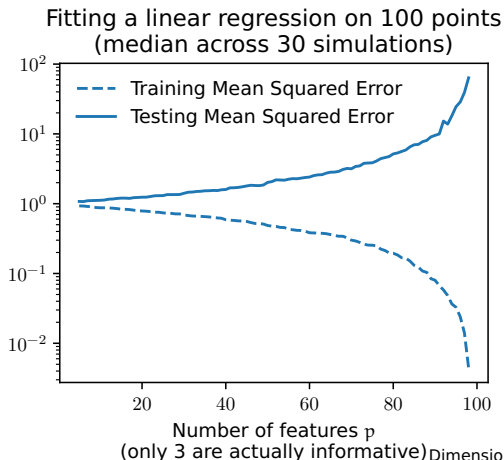
- Generate $\mathbf{X} \in \mathbb{R}^{n \times 3}$, $\boldsymbol{\beta} \in \mathbb{R}^3$, $\mathbf{e} \in \mathbb{R}^n$ and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \in \mathbb{R}^n$
- Append columns containing random noise to \mathbf{X}
- Now $\mathbf{X} \in \mathbb{R}^{n \times p}$, with $p \geq 3$, but only the first 3 columns are linked with \mathbf{y}
- Split into training and testing tests and evaluate a linear regression model: what happens when p becomes large?

See [sklearn.datasets.make_regression](#) for generating data



Model complexity: overfitting

- Model complexity increases with dimension.
- Example: a linear model in dimension p can fit exactly (0 training error) any set of $p + 1$ points.
- Risk of overfitting: fitting exactly training data but failing on test data



Univariate feature selection

- a.k.a. feature screening, filtering ...
- Check features (columns of X) one by one for association with the output y
- Keep only a fixed number or percentage of the features

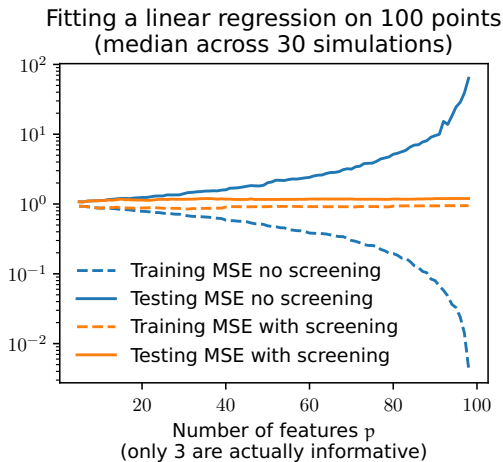
Simple (linear) association criteria

- for regression: correlation
- for classification: ANalysis Of VAriance

Read more in the scikit-learn user guide
[scikit-learn feature selection](#)

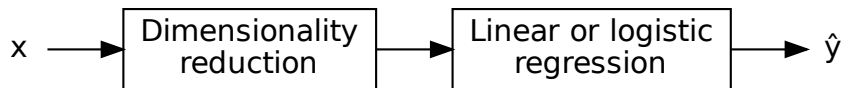
Univariate feature selection

Keeping only the 10 best features (most correlated with y)



Dataset transformations

Typical pipeline



Example



scikit-learn "transformer API": fit; transform

```
transformer = SelectKBest()  
transformer.fit(X_train)  
transformed_X = transformer.transform(X_train)
```

can also be written:

```
transformer = SelectKBest()  
transformed_X = transformer.fit_transform(X_train)
```

scikit-learn feature selection

`feature_selection.SelectKBest`

`fit:`

- compute ANOVA or correlation for each column of X
- Remember the indices of the k columns with highest scores

`transform:`

- Index input to keep only the k selected columns

`sklearn.feature_selection.SelectKBest`

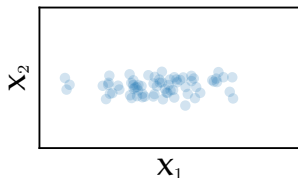
scikit-learn "transformer API": fit; transform

```
transformer = SelectKBest()  
transformed_X_train = transformer.fit_transform(X_train)  
  
transformed_X_test = transformer.transform(X_test)
```

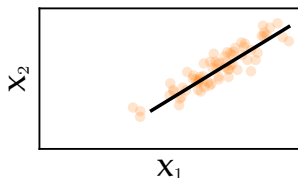
ex_04_feature_selection.py

Solution 2: linear decomposition methods

Maybe OK to drop X_2 :



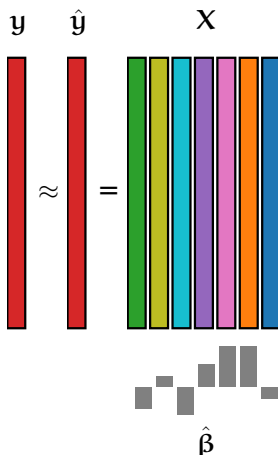
Data low-dimensional but no feature can be dropped:



Find a better referential in which to represent the data

Linear regression: projection on the column space of X

$$\hat{y} = X \hat{\beta} \quad (8)$$

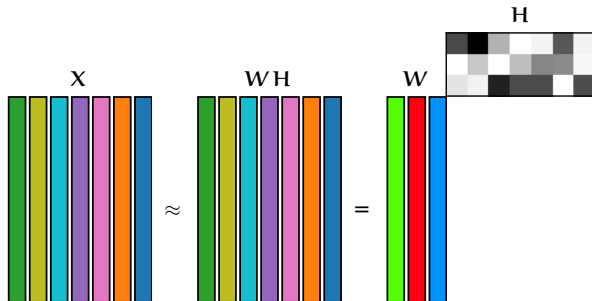


- Too many features: high variance & unstable solution
- Feature selection: drop some columns of X
- Other ways to build a family of k vectors on which to regress y ?

Linear decomposition: low-rank approximation of \mathbf{X}

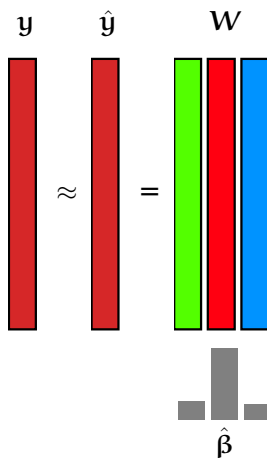
Minimize

$$\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 = \sum_{i,j} (\mathbf{X}_{i,j} - (\mathbf{W}\mathbf{H})_{i,j})^2 \quad (9)$$



Linear regression after dimensionality reduction

$$\hat{y} = W \hat{\beta} \quad (10)$$



Prediction for a new data point $\mathbf{x} \in \mathbb{R}^p$

- Find the combination of rows of \mathbf{H} that is closest to \mathbf{x} :
regress \mathbf{x} on \mathbf{H}^T
- Multiply by $\hat{\boldsymbol{\beta}}$

$$\mathbf{x} \in \mathbb{R}^p \rightarrow \text{projection} \rightarrow \mathbf{w} \in \mathbb{R}^k \rightarrow \langle \cdot, \hat{\boldsymbol{\beta}} \rangle \rightarrow \hat{y} \in \mathbb{R} \quad (11)$$

Principal Component Analysis

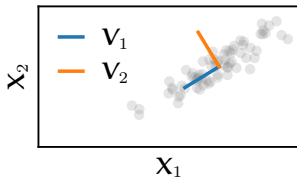
- Singular Value Decomposition of X :

$$X = U S V^T \quad (12)$$

with $X \in \mathbb{R}^{n \times p}$, $U \in \mathbb{R}^{n \times r}$, $S \in \mathbb{R}^{r \times r}$, $V \in \mathbb{R}^{r \times p}$

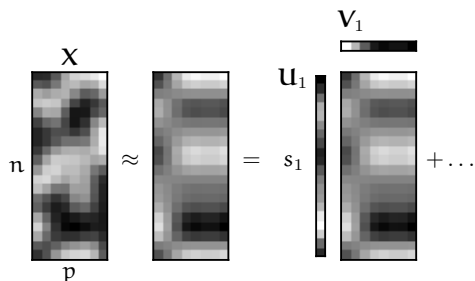
- $r = \min(n, p)$
- $S \succeq 0$ diagonal with decreasing values s_j along the diagonal
- $U^T U = I_r$
- $V^T V = I_r$

Truncating the SVD to keep only the first k components gives the best rank- k approximation of X



Singular Value Decomposition

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T \quad (13)$$



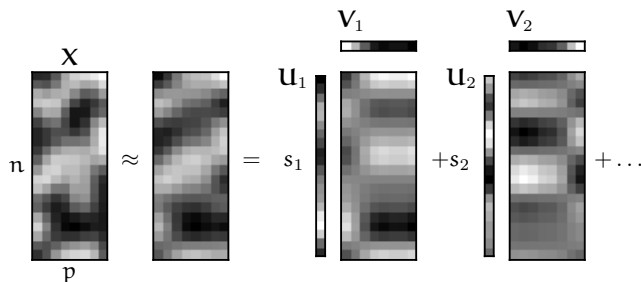
Explained variance: 0.53

$$\mathbf{U}^T \mathbf{U} = \mathbf{I}_p \quad (14)$$

$$\mathbf{V}^T \mathbf{V} = \mathbf{I}_p \quad (15)$$

Singular Value Decomposition

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T \quad (16)$$



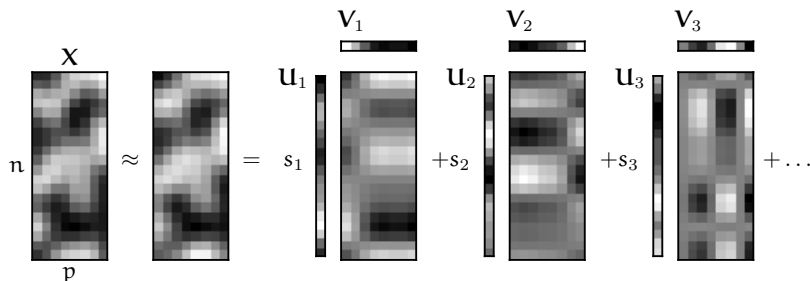
Explained variance: 0.84

$$\mathbf{U}^T \mathbf{U} = \mathbf{I}_p \quad (17)$$

$$\mathbf{V}^T \mathbf{V} = \mathbf{I}_p \quad (18)$$

Singular Value Decomposition

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T \quad (19)$$



Explained variance: 0.97

$$\mathbf{U}^T \mathbf{U} = \mathbf{I}_p \quad (20)$$

$$\mathbf{V}^T \mathbf{V} = \mathbf{I}_p \quad (21)$$

Other decomposition methods

Many other methods use the same objective (sum of squared reconstruction errors), but add penalties or constraints on the factors

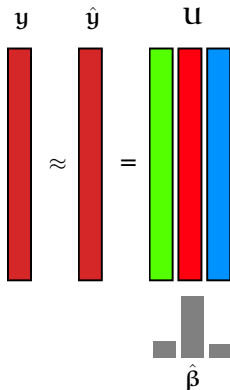
- Dictionary Learning
- Non-negative Matrix Factorization
- K-means clustering
- ...

What about y ?

- PCA is an example of *unsupervised* learning: it does not use y
- Some other methods take it into account: e.g. Partial Least Squares

Ridge regression and PCA

- Both ridge regression and PC regression compute the coordinates of y in the basis given by the SVD of X
- Ridge shrinks the coordinate along U_j by a factor $s_j^2/(s_j^2 + \alpha)$
- PC regression sets the coordinates to 0 except for those corresponding to the k largest s_j : shrinks by a factor $1_{\{j \leq k\}}$



Outline

Introduction: cross-validation

Model and hyperparameter selection

Dimensionality reduction

Conclusion

Some pitfalls with cross-validation

Overfitting the hyperparameters

- select hyperparameters with nested CV
`sklearn.model_selection.GridSearchCV`

Fitting part of the pipeline on the whole dataset

- use `sklearn.pipeline.Pipeline`

Ignoring dependencies between samples

- e.g. time series: use appropriate `cross-validation iterator`

Ignoring dependencies between CV scores

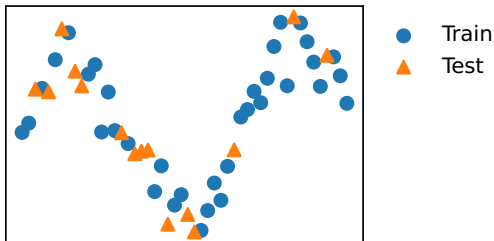
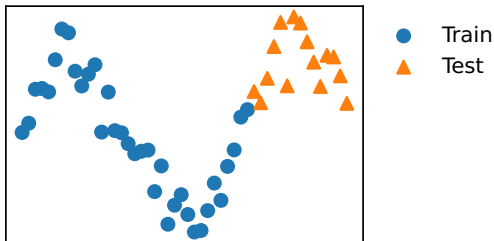
- Training sets overlap: cross-validation scores of different splits are not independent

Over-interpreting good CV scores

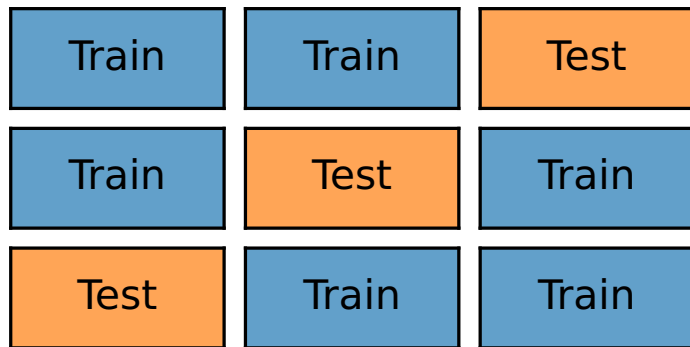
- Good CV scores do not mean the model will always perform well on a new dataset

Split choice example: time series

Which is easier?



Remember that CV training sets overlap



So the scores are not independent! Their variance can be underestimated.