

Data Science Tips

Coursework tips

you MUST clean and re-run the whole notebook before submission

- concerning randomness, please use random seed

MUST check and clean the data

provide enough explanations, in texts and equations

- explain what each function is doing
- explain each step of data handling
 - why standardise / why not standardise

Plots must have high quality

- clear titles, readable text sizes
- give axes names
- suitable scale chosen

stopping criteria is required for iterative algorithms on top of maximum iteration limit

Concisely present output for each function:

For every iterative algorithm, you MUST show convergence

- if you can, plot the cost of every step, or plot the evolution of parameters during iterations.

plot regression line to demonstrate regression result is correct

presentation must be concise and clear

- e.g. do not blindly print all the outputs, use plots or tables when necessary

Grid-search

for search of k for clustering: search one by one, do not skip e.g. search k = 1, 2, 3, ...

need to explain choice of grid-search when searching for hyper-parameters

- optimal parameter not on or close to boundary

Coursework Checklist

- ☐ check every figure has title and axes labels
- ☐ check every section has a brief introduction

- check output of every function is printed
- check every python function has docstring, and you commented as many codes as you can
- check you have explained every result in mathematical terms
- check results of different methods are summarised in a table or figure.

Basics

EDA/IDA:

- declaration of predictor space, outcome space (continuous/discrete, range etc.)

data -> Training set / test set

- avoid over-fitting. Both mean sample loss should be small

enhance robustness and ensure the model is general enough:

data -> validation set / train set

- predict outcomes on validation set
- for various choices of hyper-parameters

Do not standardise target variable y , otherwise the data loses interpretability

Linear model — explicit solutions

○ $\nabla_x(Ax)$ is not A ! It is A^T

○ similarly, $\nabla_x(y^T x)$ is y

○ Possible problems with matrices:

○ large condition number of $X^T X$ (or if not invertible, infinite condition number)

- this makes the variance of estimator β^* large

works even for large p : shrinkage methods

- Ridge Regression: add penalty for $\|\beta\|$, by adding $\lambda\|\beta\|$ behind the goal function

T-fold cross validation:

- $T = 5$ for small data sets, $T = 10$ for larger data sets.

remember to update intercept in regression.

dealing with minority classification problems

proportion of TP/ TN is small

method 1: use bootstrap with weights to scale up the minority class, scale down the majority class

method 2: use weighted loss function.

$$\text{Example: cross-entropy} \quad E = \sum_{\mathbf{x} \in \Omega} w(\mathbf{x}) \log(p_{\ell(\mathbf{x})}(\mathbf{x}))$$

Contingency tables

{0, 1} binary case

Precision = TP / (TP + FP)

- how much we can trust the model when it predicts positives

recall = TP / (TP + FN)

- proportion of positives correctly identified

SVM

Intercept of SVM should not be optimised via gradient descent, but use the formulae

$$\mathbf{w}^* = \alpha_+ \mathbf{x}_+ - \alpha_- \mathbf{x}_- \\ b = 1 - \mathbf{x}_+ \cdot [\alpha_+ \mathbf{x}_+ - \alpha_- \mathbf{x}_-] = 1 - \alpha_+ \underbrace{\mathbf{x}_+ \cdot \mathbf{x}_+}_{\star} + \alpha_- \underbrace{\mathbf{x}_+ \cdot \mathbf{x}_-}_{\star}$$

to update

for more assistance on the course material, see the pdf file *Data Science Auxiliary Notes* in the folder