ML - Week 5 - Model Validation I

▷ Data Collection & annotation

Problem understanding
- what is the real world issue?
- what is the role of ML - is it needed & How
- Stakeholders
- risks & issues

Data Sources in General:
1. collected : e.g. sensors
2. internet : scraping
3. crowdsourcing: encourage ppl to contribute

Data annotation for obtaining the labels:
1. Self reporting / tagging
2. Expert annotators
3. crowdsource - captcha

▷ Data Prep & Preprocessing

→ Data extraction

- @ collection
- Post collection
- Domain specific preprop

Example → magnitude (pos & neg) from Neg/Pos Data

Exp/ UK Metrological data - 2000 to 2022 - From dabs
- remove missing values

Input Scaling ⟶ Normalization

- bring data into same distribution
- Dists might be wide vs narrow

Standard Scaling = Scales to mean $\mu = 0$ & SD $\sigma = 1$

$$\check{x}^n_d = \frac{x^n_d - \mu_d}{\sigma_d}$$

Min-Max Scale = min = 0 , max = 1

$$\check{x}^n_d = \frac{x^n_d - x^{min}_d}{x^{max}_d - x^{min}_d}$$

remove Missing

- Dont need to Normalize if using Dec trees
- Otherwise always normalize

Missing Data

- Equipment ~~error~~ issue
- Human ~~error~~ factor

Options:

- Discard instances: easy, but effects prediction
- **Imputation** – replace data, Many methods
- Surrogacy – use whatever feature are available

Missing Output labels

- Discard
- Self-supervision ⟶ make use of available data
- unsupervised learning

Data Augmentation

DA is the application of random transformations to training data

this is to force ML model to learn informative Patterns & enhance ability to generalize

Rather than memorize Train data

Example → images, h x w dims, rot, flip, translate, shear, crop
- force to learn patterns, not images

label imbalance

undersample maj class - randomly

Oversample Min class - match maj class

oversample Min class w/ data augmentation

weights in the loss function - Penalize more if Min class is incorrect
- weighted MSE class k_i: