Practical Issues in Machine Learning

NLP II 2025

Assoc. Prof. Attapol Thamrongrattanarit

Machine Learning so far

- Naive Bayes: representation, learning, and prediction
- Logistic Regression: representation, learning, and prediction
- Linguistic theory for a specific task for sentiment analysis

Practical Issues

- Multi-label vs Multi-class
- Overfitting and underfitting: Regularization

Practical Issues

- Multi-label vs Multi-class
- Overfitting and underfitting: Regularization

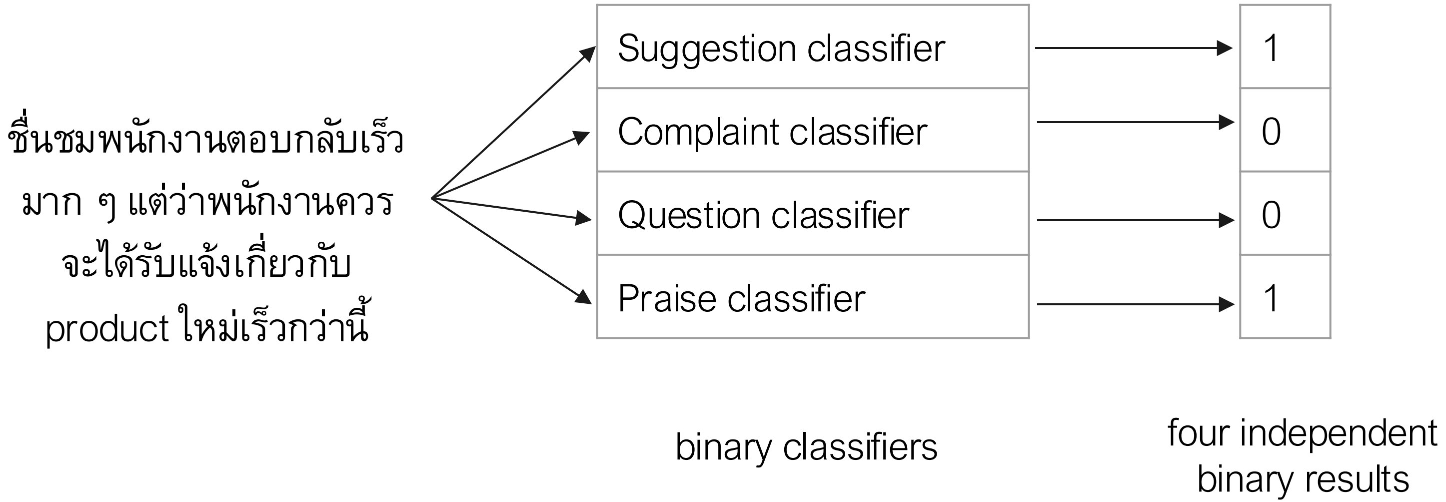
Multi-class vs multi-label

- Multi-class classification assumes that we label text input with one out of the pre-defined k classes (labels or categories). Also called k-way classification
  - Example?
- Multi-label classification assumes that we label text input with at least one of the pre-defined k classes.
  - Each class is not mutually exclusive. Example?
  - There can be multiple answers.

Example

- Book genre classification
- Product classification
- Social media content classification
- News article classification
- Customer feedback classification

How to deal with multi-label classification



Multi-class (k-way classification) vs Multi-label

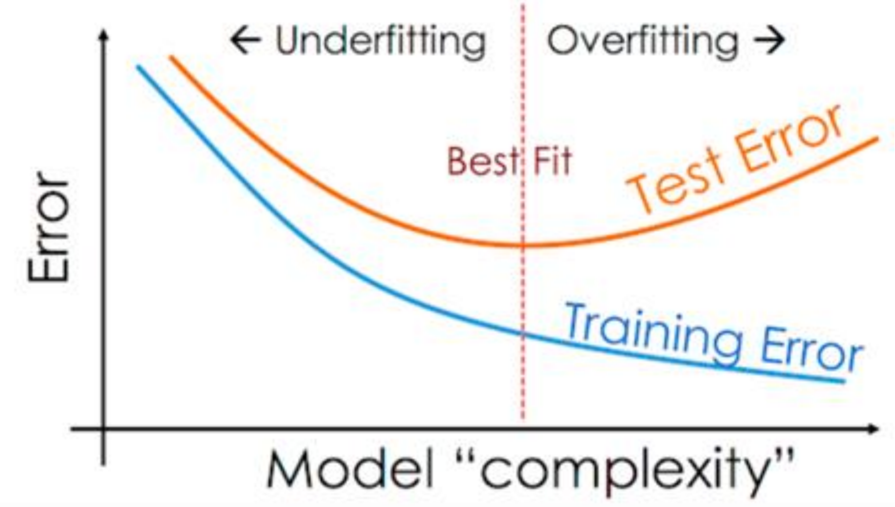
- Multi-class classification is done with one single model. But we have to simplify our analysis to one class per text input and design the label set to be mutually exclusive (เป็นอันนึงแล้วห้ามเป็นอีกอันนึง).
- Multi-label classification is done with k models. We have to train k models on the same dataset. This is a more complex solution. The choice is up to the use case.

Practical Issues

- Multi-label vs Multi-class
- Overfitting and underfitting: Regularization

The Tension

Adding more features (more complexity) reduces the test error to a certain point and then increases it.



Consider these features: non-conflicting features

Text	**Label (Y)**	predictable	and	boring	very	few	laughs	short	but	powerful	fun	good
predictable and boring	negative	**1**	**1**	**1**								
very few laughs	negative				**1**	**1**	**1**					
short but boring	negative			**1**				**1**	**1**			
very very powerful	positive				**2**					**1**		
fun and good good laughs	positive		**1**				**1**				**1**	**2**

Overfitting

Text	**Label (Y)**	predictable	and	boring	very	few	laughs	short	but	powerful	fun	good
predictable and boring	negative	1	1	1								
very few laughs	negative				1	1	1					
short but boring	negative			1				1	1			
very very powerful	positive				2					1		
fun and good good laughs	positive		1				1				1	2

- Since there is none (or too little) contradicting evidence from the dataset, the loss function decreases if we make the weights for these non-conflicting features really really big. Consider "boring but many laughs"

Overfitting

- Overfitting refers to the situation where the model fits too closely to the training data, and as a result, it doesn't generalize to unseen data. This happens when we have a lot of features (e.g. bag-of-word features)
- Underfitting refers to the situation where the model is too simple to fit well to the training data, and as a result, it performs poorly on unseen data. This happens when we have too few features or the features are ineffective.

Regularization

- Regularization is a technique to mitigate the problem of overfitting.
- One of the most popular ways is L2-regularization where we have to 'pay' for the weights to make the model less confident about any single feature. The larger the weights, the more expensive.

Consider these features: non-conflicting features

Text	**Label (Y)**	predictable	and	boring	very	few	laughs	short	but	powerful	fun	good
predictable and boring	negative	**1**	**1**	**1**								
very few laughs	negative				**1**	**1**	**1**					
short but boring	negative			**1**				**1**	**1**			
very very powerful	positive				**2**					**1**		
fun and good good laughs	positive		**1**				**1**				**1**	**2**

L2 Regularization

$$L_{CE}(W,b) = -\frac{1}{N}\sum_{i=1}^{N} \log P(Y = y^i | x^i, W, b)$$

$$L_{CE+L2}(W,b) = -\frac{1}{N}\sum_{i=1}^{N} \log P(Y = y^i | x^i, W, b) + \alpha((\sum_i \sum_j w_{ij}^2) + \sum_i b_i^2)$$

Weight matrix (W)	bias	'against'	'love'	text length
positive	0.15	-2	-1	0.0004
negative	-0.2	2	-0.2	0.005
neutral	1	-1	0.4	-0.00001