writeup

Gasper Qian

21/04/2022

Deliverable 8.1 (4 points):

Briefly describe your bakeoff design.

The training data is pre-processed with stem and stopwords which replaced morphological variants of the root words with the root words itself, and remove stopwords (common words for connecting sentences and can be ignored) to concentrate on a smaller size of data with more obvious and represented features. This contributes to a 1% increase in accuracy compare to the original training set.

Logistic regression is used with Neg Log-likelihood loss function on our modelling. L2 regularization is applied under the optimizer, which helped on optimizing the loss function with penalizations. This contributes to a 1% increase in accuracy compare to the original training set.

Other loss functions are attempted, but did not seem to give a better accuracy on validation set.

Deliverable 8.2 (4 points):

Why is Naive Bayes a generative model? Why is Perceptron a discriminative model? Which family does logistic regression belong to? What do you see as a core difference between a generative model and a discriminative model?

Naive Bayes is a generative model because we try to find the argument that maximize the joint distribution, which can be broken down to the prior times the likelihood, i.e. $P(X, Y) = P(Y)P(X|Y)$.

Perceptron a discriminative model because they are trained directly to classify given X to maximize the conditional probabilities P(Y|X), and cannot be used to estimate the probability of X or generate X | Y.

A core difference between a generative model and a discriminative model is that one tries to maximize the joint distribution P(X,Y), the other tries to maximize the conditional distribution P(Y|X).

Deliverable 8.3b (4 points):

Question: What kind of prior $P(\theta)$ leads us to the solution with Laplace smoothing? Why? (Hint: a kind of distribution mentioned in Eisenstein Ch. 5.5.1 could be helpful)

Dirichlet distribution is a typical choice for priors that leads us to the solution with Laplace smoothing and MAP, especially for the multinomial and categorical parameters. This is because it defines a probability

on exactly the set of vectors that can be parameters: vectors that sum to one and include only non-negative numbers, which fits the scenarios of most of θ conditions, e.g. in project 1, θ is represented by a set of labels, which is multinomial with probabilities sum up to 1.