Reflection Week 3

Summary

This week's lectures covered classification models, including binary classification models and multiclass classification, probability concepts used in calculating weights for classifying, as well as miscellaneous concepts like decision boundaries, accuracy, precision, errors, and logistic models.

Concepts

- Classification model: a model that reads some input and generate an output that classfies the input into some category (class)
- Decision boundary: a region of a problem space between two or more class (label) regions in which the output label of a classfier is ambiguous
- Classification error $= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{y_i \neq \hat{y}_i\}$
- Accuracy = 1 - error, should be at least 1/k with k = number of classes
- Confusion matrix:
  - True positive
  - True negative
  - False negative (predicted false but should be true)
  - False positive (predicted true but should be false)
- Threshold works based on this simplified execution:

```
if score(x) > alpha:
    predict y = +1
else:
    predict y = -1
```

- If I never want to make a false positive prediction (always have $\hat{y} = -1$) then let $\alpha = \infty$ so that the evaluation always goes into the else statement . On the other hand, if I never want to make a false negative prediction, let $\alpha = -\infty$ so it never makes it to the else statement
- Probability classifiers generate a probability of one input x being in a label y, which improves interpretability.
- Logistic regression: a type of regression using the sigmoid curve that generates a probability score that assigns some probability (0 to 1) to the input belonging to a category.
  $sigmoid(Score(x)) = \frac{1}{1+e^{-Score(x)}} = P(y_i = +1|x_i, w)$ with w as the vector of weights of the model
- Finding weights by maximizing the log-likelihood by using gradient ascent performing on models graphed using the function: $\hat{w} = argmax_w \ell(w) = argmax_w \sum_{i=1}^{n} \log(P(y_i|x_i, w))$
- Again, to prevent overfitting, use regularization and optimize the number of features used. By increasing the number of features, models become more complex, coefficients get larger in magnitude, but probabilities get closer to 0 and 1, narrowing down the arbitrary area of the decision boundary.

Concerns

- What does it mean when we say that we choose the weights that maximize the probabilities?
- How do we define scores for multiclass models?
- What is an acceptable size for the arbitrary zone of the decision boundary?