1 Probability & Statistics

$P(A|B) = \frac{P(A \cap B)}{P(B)}$
Probability of A given B is the joint probability of A and B divided by B.

$P(A \cap B) = P(A|B) \cdot P(B)$
The Chain Rule, inferrable from above.

1.1 Bayes' Rule

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
$P(A|B)$ is the posterior probability, informed by the information given by the event B occuring. $P(A)$ is the prior probability.

We can also say that A is our hypothesis and B is our evidence. If I am looking for the chance that I have cancer, with no test it is A but with a test it becomes $P(A|B)$ where B is the outcome of the test.

In the case of a test where a positive result is B, Bayes' rule can also be thought of as
$\frac{TP}{TP+FP} = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|!A)P(!A)} = \frac{P(B|A)P(A)}{P(B)}$

1.2 P-Value

2 Information Theory

TODO: entropy, cross entropy, perplexity

3 Vocabulary

1. **Precision** is $\frac{TP}{TP+FP}$, or "how many things you marked positive were actually positive".

2. **Recall** is $\frac{TP}{TP+FN}$, or "how many of the actually positive things did you mark positive".

3. **Accuracy** is $\frac{TP+TN}{TP+TN+FP+FN}$, or "how many of the total predictions did you get right".

4. **F1 Score** is $(1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision)+recall}$ where recall is β times as important as precision. Often $\beta = 1$ for $2 \cdot \frac{precision \cdot recall}{precision+recall}$

   Precision, Recall and F1 Score all focus on your true positives. F1 is less useful if you need to know about your true negatives.

5. **Type 1 Error** is a false positive, **Type 2 Error** is a false negative.

6. **Bias** is how dissimilar you are to the distribution of the training data, either for reasons of intentional bias introduction or because of a simple model.

7. **Variance** is how similiar your model distribution is to the training data, which requires that your model is sufficiently complex and models the training data closely.

8. **Overfit/Underfit** is building a distribution too similar or different to the training data, respectively. Too much bias = underfit. Too much variance = overfit.

4 Preventing Overfitting

k-fold cross validation, l1 and l2 regularization

4.1 Neural Methods

l2 reg common dropout also common batch norm??

5 Decision Trees

6 Naive Bayes

multinomial model, binomial model

7 Regression

logistic regression vs binomial regression vs linear regression

8 Support Vector Machines

9 Neural Netorks

backprop

9.1 Feed-Forward

Commonly ”multi layer perceptron” for classification

9.2 Long Short-Term Memory

9.3 Transformers

10 LDA

11 Algorithms

11.1 Bach Gradient Descent

11.2 Stochastic Gradient Descent

11.3 Beam Search

11.4 TF-IDF