Hw2

(1)

$$P(W = T \mid Class = T) = 1/2$$

$$P(W = T \mid Class = F) = 2/3$$

$$P(X = T \mid Class = T) = 1/2$$

$$P(X = T \mid Class = F) = 1/3$$

$$P(Y = F \mid Class = T) = 1/2$$

$$P(Y = F \mid Class = F) = 1/3$$

$$P(Class = T) = 2/5$$

$$P(Class = F) = 3/5$$

$$\therefore \quad 1/2 * 1/2 * 1/2 * 2/5 = \frac{1}{20}$$

$$>$$

$$2/3 * 1/3 * 1/3 * 3/5 = 2/45$$

$\therefore$ Class predicted is T

(b). Only the prior changes.

$$\therefore \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{102}{105}$$

$$> \frac{2}{3} * \frac{1}{3} * \frac{1}{3} * \frac{3}{105}$$

$\therefore$ Prediction ~~is~~ ~~F~~ remains true

(c) Yes, Since Naive Bayes ~~just~~ uses ~~[illegible]~~ probabilities ~~data~~ it does not matter if data is noisy or not. For the same features with different labels, their conditional probabilities would change to reflect the presence of noise

2. Ideally no Regularization ~~[illegible]~~ tends to change the features to avoid over-fitting and should not ideally affect expresiveness of the model.

3. For a 1-NN, since every training example is a nearest neighbor of itself. accuracy is 100%.

For a 3-NN, 100% accuracy is not guaranteed.

e.g.

the "A" datapoint is labeled wrongly using 3-NN

4. Bayesian learning learns a distribution over the parameters. using a prior.

MAP learning, approximates the parameter distributions in Bayesian learning. to a single parameters which corresponds to the highest-probability values in the distributions.

Max-likelihood. estimates a single parameter without worrying about the prior, and using only the data. to determine the optimal parameters that best fit the data.

5. Sample complexity is ~~inversely~~ ~~inversely~~ proportional to VC-Dimension. Since 1-NN has a much larger VC-Dimension than logistic regression (linear decision boundries), learning ~~it~~ ~~is~~ 1-NN is harder than logistic regression.

6. VC-Dimension = $(K+1)$

$\epsilon = 0.01$

$\delta = 0.05$

$$\therefore m \geq \frac{1}{0.01}\left[4 \log_2\left(2/0.05\right) + 8\left[K+1\right]\log_2\left(13/0.01\right)\right]$$

7.

We need to estimate the size of the hypothesis class.

$$|H| = 2^8 \cdot 2^8 \cdot 2^8 \cdot 2^8 = 2^{32}$$

$$\epsilon = 0.01$$

$$\delta = 0.1$$

$$m \geq \frac{1}{0.01}\left[\ln(2^{32}) + \ln(1/0.1)\right]$$

$$\geq \frac{1}{0.01}\left[32 \ln 2 + \ln(1/0.1)\right]$$

$$=$$

8. For a finite concept class H

$$VC(H) \leq \log_2 |H|$$

$\therefore$ using $VC(H)$ is better than using $|H|$ which would give us a tighter sample complexity.

If $d_1 < d_2$, this only means that ~~in~~ the worst-case sample complexity of L_2 is worse than L_1. However, on specific datasets L_2 may outperform L_1.

E.g. 1-NN can perform better than logistic regression ~~on some~~ for specific datasets even though 1-NN has a larger VC-dimension than logistic regression.