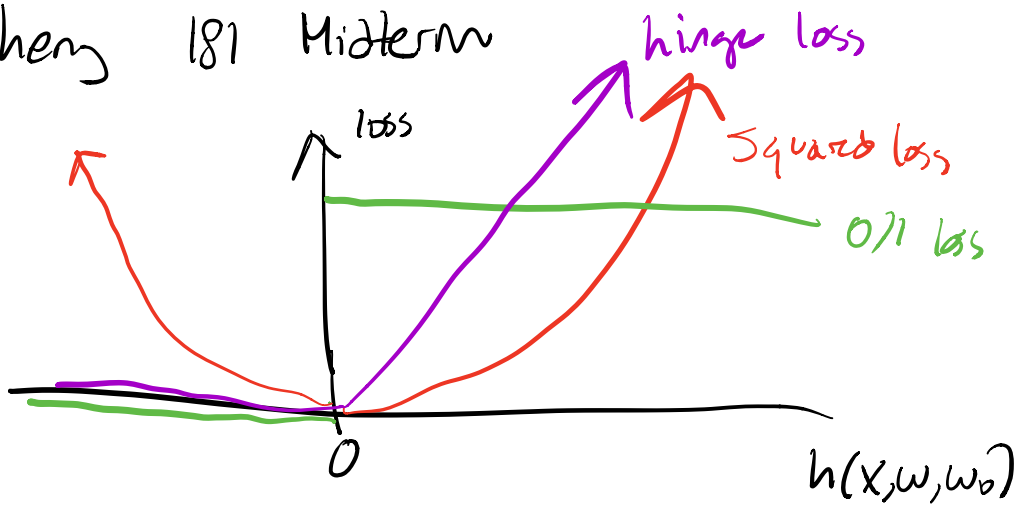
Chao Cheng 181 Midterm

1a.



b. 0/1 → B

hinge → A

Squared → C

c. L1 loss, as it is less sensitive to outliers and can be used in gradient descent because it is differentiable.

d. 0-1 loss: $x = (-1, 1)$

hinge loss: $x = (-1, 1)$

e. 0-1 loss: $x = (-1, 1)$

hinge loss: $x = (1, 2)$

2. a. From Bayes Rule, we have

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)} \propto P(D|w)P(w)$$

since $P(D)$ is constant, this implies

$$w_{MAP} = \underset{w}{\operatorname{argmax}}\ P(w|D) = \underset{w}{\operatorname{argmax}}\ P(D|w)P(w)$$

b.

$$w_{MLE} = \underset{w}{\operatorname{argmin}}\ -\log P(D|w)$$

$$w_{MAP} = \underset{w}{\operatorname{argmin}}\ -\log P(D|w) - \log P(w)$$

The w_{MAP} has an extra $-\log P(w)$ term which serves as a "regularization" tool by taking into account our knowledge of the prior, thus making it less susceptible to overfitting.

c. when $P(w) = 1$ (ie we know for certain what w is)

d. $N \to \infty \Rightarrow w_{MAP} \to w_{MLE}$

which makes sense as the more data we have, the two estimators converge

$s_0 \to \infty \Rightarrow \mu_{map} \to \mu_{mle}$

the larger the prior variance, the more the two estimators converge

$\sigma^2 \to \infty \Rightarrow \mu_{map} \to m_0$

the greater the variance, the less certain we are about the data, so the μ_{MAP} is just the prior mean

3.a. θ = prior probability of student being CS

$\pi_{0,d}$ = probability of student being Stat if the d^{th} feature = 1

$\pi_{1,d}$ = probability of student being CS if d^{th} feature = 1

1. generative: we fit a probability distribution to the data which can then be used to "generate" more data in the future

naive: we assume the probabilities of each feature being 0 or 1 are independent (thus we can take

the simple product of probabilities)

c. From Bayes Rule

$$P(y=1 \mid x) \propto P(x \mid y=1) P(y=1)$$
$$P(y=0 \mid 0) \propto P(x \mid y=0) P(y=0)$$

we want to find the higher of these two

which is equivalent to finding the higher of these two expressions

thus whichever is larger (ie. $a-b>0 \Rightarrow a>b$ & $a-b<0 \Rightarrow a<b$) is the class we should classify x as.

d.

$$\left(x^T \ln \pi_1 + (1-x)^T \ln(1-\pi_1)\right)(\theta)$$
$$-\left(x^T \ln \pi_0 + (1-x)^T \ln(1-\pi_0)\right)(1-\theta)$$

$$= x^T\left(\theta \ln \pi_1 - (1-\theta)\ln \pi_0\right) + (1-x)^T\left(\theta \ln(1-\pi_1) - (1-\theta)\ln(1-\pi_0)\right)$$

e. if $\pi_{1,d} > \pi_{0,d}$ for a feature d, then the expression will have a higher weight for that feature which will skew it to be >0

f. if $\pi_{1,d} = \pi_{0,d}$ the two weights cancel out each other so there is no net effect on the expression

/expected val.

4. a. $\bar{y}$ is the target mean, ie mean of $y|x$ dist.

$\bar{f}(x)$ is the mean of our predictions

$(\bar{y} - \bar{f(x)})^2$ captures the "squared error" of our prediction, ie, how far we are off by in our prediction squared, and we want to minimize this to get the best fitting prediction function f

b. the variance is simply defined as the expected value of the square of the difference between each individual prediction $f(x_i)$ and the mean $\bar{f}(x)$

it measures the "variability" / spread of our predictions

c. complex models fit the data better and thus have higher variance. By bias-variance tradeoff, simple models have less variance but higher bias, ie. they don't fit the data as well and therefore doesn't respond to new training data as much, therefore having higher bias.

d. explained above, but complex models attempt to fit the data more which leads to greater variance.

e. with more data, the variance of complex models can be decreased, but if we don't have lots of data, it's better to use a simple model with low

Variance but high bias.

f. Cross-validation.

5.

a. $$P(\{y_n\}|\{x_n\},w) = \prod_{i=1}^{n} \binom{3}{k} \theta_n^{y_i} (1-\theta_n)^{3-y_i}$$

b. $$\log P(\{y_n\}|\{x_n\},w) = n \log\binom{3}{k} + \sum_{i=1}^{n} y_i \log\theta_n + (3-y_i)\log(1-\theta_n)$$

$$L = -\log P = -n\log\binom{3}{k} - \sum_{i=1}^{n} y_i \log\theta_n - \sum_{i=1}^{n} (3-y_i)\log(1-\theta_n)$$

c.

$$\frac{dL}{dw} = \frac{dL}{d\theta} \cdot \frac{d\theta_n}{dw}$$

$$= \left(-\frac{\sum y_i}{\theta} + \frac{\sum(3-y_i)}{1-\theta}\right)\left(\sigma(w^T x_n)(1-\sigma(w^T x_n))\right)(x)$$