WRITTEN EXERCISES

DATE:

(1) 80 Cornell students
30 of them are Master's
50 of them are Phd

Master's students (30 out of 80):
8 Master's students who bike
9 Master's students who ski

PhD Students (50 out of 80)
30 PhD students who bike
12 PhD students who ski

TARGET: $y = 1$ if student is PhD
$y = 0$ if student is Master's.

Features: x_1 = a binary indicator of whether a student ~~somebody~~ bikes
x_2 = a binary indicator of whether a student skis.

~~(a)~~ $P(x_1) = \frac{8+30}{80} = \frac{38}{80}$

$P(x_2) = \frac{9+12}{80} = \frac{21}{80}$

$P(Y = y \mid X) = (x_1, x_2))$

$$P(Y \mid X) = \frac{P(X \mid Y)\,P(Y)}{P(X)}$$

$\cancel{\frac{P(x_1 = 0)}{Y = 0}}$ $P(x_1 = 0 \mid Y = 0) = \frac{22}{30}$ $P(x_2 = 0 \mid Y = 0) = \frac{21}{30}$

$\cancel{P(x_2}$ $P(x_1 = 0 \mid Y = 1) = \frac{20}{50}$ $P(x_2 = 0 \mid Y = 1) = \frac{38}{50}$

$P(Y = 0) = \frac{30}{80}$ $P(Y = 1) = \frac{50}{80}$

DATE:

a) Naive Bayes classifier assumes that predictors/ features in a Naive Bayes model are conditionally independent, or unrelated to any of the other feature in a model. It also assumes that all features contribute equally to the outcome. In this context, the features of whether a student bikes (x_1) or whether a student skis (x_2) are conditionally independent i.e. observing whether a student bikes has no influence on observing whether a student skis. Both features (whether a student bikes or skis) contribute equally in determining to which class a student belongs (Master's or PhD).

b) $P(Y=0 \mid x_1=0; x_2=0) = \dfrac{P((x_1=0, x_2=0) \mid Y=0) \cdot P(Y=0)}{\text{Total probability of } P((x_1=0; x_2=0))}$

$$= \frac{P((X_1=0, X_2=0) \mid Y=0) \cdot P(Y=0)}{P((X_1=0, X_2=0) \mid Y=0) \cdot P(Y=0) + P((X_1=0, X_2=0) \mid Y=1) \cdot P(Y=1)}$$

$$= \frac{\frac{\cancel{22}^{11}}{\cancel{30}} * \cancel{3} \; \frac{\cancel{21}^{7}}{\cancel{30}\,\cancel{10}\,5} * \frac{\cancel{30}}{80}}{\left(\frac{22}{30} * \frac{21}{30} * \frac{30}{80}\right) + \left(\frac{38}{50} * \frac{20}{50} * \frac{50}{80}\right)}$$

$$= \frac{77/400}{\frac{77}{400} + \frac{19}{100}}$$

$$= \frac{77}{153}.$$

c) Given that every PhD who skis also bikes.

$P(B \cap S)$ for PhD $= 12$

$P(X_1=0, X_2=0 \mid Y=1) = 1 - \frac{30}{50} = \frac{20}{50}$.

Using the same formula as above (1b)

$$P(Y=0 \mid X_1=0; X_2=0) = \frac{P((X_1=0, X_2=0) \mid Y=0) \cdot P(Y=0)}{P((X_1=0, X_2=0) \mid Y=0) \cdot P(Y=0) + P((X_1=0, X_2=0) \mid Y=1) \cdot P(Y=1)}$$

$$= \frac{77/400}{\frac{77}{400} + \left(\frac{\cancel{20}^{1}}{\cancel{50}} * \frac{\cancel{50}}{\cancel{80}\,4}\right)}$$

$$= \frac{77/400}{77/400 + \frac{100}{400}} = \frac{77}{177}$$

DATE

$P_\theta(y=k) = \phi_k$.

2a) Log Likelihood estimate.

$$\log P_\theta(x^i, y^i)$$

$$P_\theta(y) = \phi_k.$$

$$P_\theta(x|y) = \psi_{jkl}$$

$$P_\theta(x|y=k) = \prod_{j=1}^{d} P_\theta(x_j|y=k)$$

$$P(x^i, y^i) = \log P_\theta(x^i, y^i)$$

$$= \log \prod_{j=1}^{d} P(x_j, y=k)$$

Now we know,

$$P + (1-P) = 1.$$

Bernoulli's distribution states

$$P(x) = p^x(1-p)^{1-x}.$$

$$P(x_j, y=k)$$

Now, Bernoulli distribution has a single parameter P, We replace y with $P_\theta(y=k) = \phi_k$.

The question gives us a distribution over a vector of features x, ~~can be~~ which is.

$$P_\theta(x|y=k) = \prod_{j=1}^{d} P_\theta(x_j|y=k).$$

We factor in the probability function above with

$$\phi_k^{x_j}(1-\phi_k)^{1-x_j}.$$

Thus, we get $\prod_{j=1}^{d} \phi_k^{x_j}(1-\phi_k)^{1-x_j}$

When we add log to both sides of the equation, we get :

$$\log \prod_{j=1}^{d} \phi_K^{x_j} (1-\phi_K)^{1-x_j}$$

Due to the log operators the multiplication is correlated to summation.

$$\log \prod_{j=1}^{d} \phi_K^{x_j} (1-\phi_K)^{1-x_j} = \sum_{j=1}^{d} \log \phi_K^{x_j} + \sum_{j=1}^{d} \log (1-\phi_K)^{1-x_j}$$

~~Using rules of log power.~~

Σ We replace j with i and d with n.

$$\log \prod_{i=1}^{n} \phi_K^{x_j} (1-\phi_K)^{1-x_j} = \sum_{i=1}^{n} \log \phi_K^{x_i} + \sum_{j=1}^{n} \log (1-\phi_K)^{1-x_i}$$

$$= \sum_{j=1}^{n} x^j \log \phi_K + \sum_{i=1}^{n} (1-x_i) \log (1-\phi_K).$$

The log elements do not depend on i/j when we bring them out of the summation.

$$\log \phi_K \sum_{i=1}^{n} x^i + \log (1-\phi_K) \sum_{i=1}^{n} (1-x_i).$$

$$= \log \phi_K \sum_{i=1}^{n} x^i + \log (1-\phi_K) (N - \sum_{i=1}^{n} x_i)$$

We take the derivative of the above at 0 to get maximum likehihood.

$$\frac{\sum_{i=1}^{n} x^i}{\phi_K^* \ln 10} - \frac{(N - \sum_{i=1}^{n} x_i)}{(1-\phi_K^*) \ln 10} = 0$$

Unifying the denominator

$$\frac{(1-\phi_K^*) \sum_{i=1}^{n} x_i - \phi_K^* (N - \sum_{i=1}^{n} x_i)}{\phi_K^* (1-\phi_K^*) \ln 10} = 0$$

DATE:

$$(1-\phi_k^*)\sum_{i=1}^{n} x^i - \phi_k^*(N-\sum_{i=1}^{n} x^i) = 0$$

$$= \sum_{i=1}^{n} x^i - \cancel{\phi_k^* \sum_{i=1}^{n} x^i} - \phi_k^* N + \cancel{\phi_k^* \sum_{i=1}^{n} x_i} = 0$$

$$\Rightarrow \phi_k^* N = \sum_{i=1}^{n} x^i$$

$$\Rightarrow \phi_k^* = \frac{\sum_{i=1}^{n} x^i}{N}$$

we know $\sum_{j=1}^{n} x^j$ is number of features for class k so we can write it as:

$$\phi_k^* = \frac{N^k}{N}.$$

b) $$\ell(\Theta) = \sum_{i=y^i=k} \log P(x_j^i \mid y^i ; \psi_{jkl})$$

$$= \sum_{i=y^{(i)}=k} \log \left(\frac{\psi_{jk} x_j^i}{\sum_{k=1}^{K} \psi_{jkl}} \right)$$

$$= \sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=y^i=k} \log \frac{\psi_{jk} \, x_j^i}{\sum_{l=1}^{L} \psi_{jkl}}$$

$$\frac{d}{d\psi_{jkl}} (\ell(\Theta)) = \frac{d}{d\psi_{jkl}} \left(\sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=y^i=k} \log \frac{\psi_{jk} x_j^i}{\sum_{l=1}^{L} \psi_{jkl}} - \right).$$

$$= \frac{d}{d\psi_{jkl}} \left(\sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=y^i=k} \log \frac{\psi_{jk} x_j^i}{\sum_{l=1}^{L} \psi_{jkl}} \right)$$

$$= \frac{d}{d\psi_{jkl}} \left(\sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=y^i=k} \left(\log(\psi_{jk} x_j^i) - \log\left(\sum_{l=1}^{L} \psi_{jkl}\right)\right)\right)$$

$$= \frac{n_{jkl}}{\psi_{jkl}} - n_k = 0$$

$$\Rightarrow \psi_{jkl}^* = \frac{n_{jkl}}{n_k}.$$