$$l(\mathbf{w}) = \sum_i \left[y^i \mathbf{w}^{\mathbf{T}} \mathbf{x}^i - \log\left(1 + \exp\left(\mathbf{w}^T \mathbf{x}^i\right)\right)\right]$$

Consider the above log-likelihood objective for logistic regression. What is $\frac{\partial l}{\partial w_k}$?

A. $\sum_i \left[y^i x_k^i - \frac{x_k^i}{1 + \exp(\mathbf{w}^T \mathbf{x}^i)} \right]$

B. $\sum_i \left[y^i x_k^i + \frac{x_k^i}{1 + \exp(-\mathbf{w}^T \mathbf{x}^i)} \right]$

C. $\sum_i \left[y^i x_k^i - \frac{\exp\left(\mathbf{w}^T \mathbf{x}^i\right) x_k^i}{1 + \exp(\mathbf{w}^T \mathbf{x}^i)} \right]$

D. $\sum_i \left[y^i x_k^i - \frac{x_k^i}{1 + \exp(-\mathbf{w}^T \mathbf{x}^i)} \right]$

Answer: C, D

Assuming that our computer has infinite numerical precision, which of the followings are true?

A. $\frac{\partial l}{\partial w_p}$ will become smaller and smaller but never reaches zero

B. $\frac{\partial l}{\partial w_p}$ will reach zero when all positive training examples are correctly classified

C. As $\frac{\partial l}{\partial w_p}$ gets smaller and smaller, w_p will approach zero.

D. As $\frac{\partial l}{\partial w_p}$ gets smaller and smaller, w_p will approach infinity

Answer: A and D

Apply Naive Bayes to this data and compute $P(y=1|1,0,0)$

X_1	X_2	X_3	Y
1	1	1	0
1	1	0	0
0	0	0	0
0	1	0	1
1	0	1	1
0	1	1	1

A. $\frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$

B. $\dfrac{\frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}}{\frac{2}{6} \times \frac{2}{6} \times \frac{3}{6}}$

C. $\dfrac{\frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} + \frac{1}{2} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3}}$

D. $\dfrac{\frac{1}{2} \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{2}{3} \times \frac{1}{2} \times \frac{2}{3}}$

E. $\frac{1}{3} \times \frac{1}{2} \times \frac{1}{3}$

Answer: C

For k-class classification with v binary features, how many parameters do we need to learn for the Naive Bayes classifier?

A. $k * (2^v - 1)$

B. $k + v - 1$

C. $k * v$

D. $k * v + k - 1$

Answer: D