DSC 40A Fall 2025 - Group Work Week 10

due Monday, Dec 1st at 11:59PM

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. **One person** from each group should submit your solutions to Gradescope and **tag all group members** so everyone gets credit.

This worksheet won't be graded on correctness, but rather on good-faith effort. Even if you don't solve any of the problems, you should include some explanation of what you thought about and discussed, so that you can get credit for spending time on the assignment.

In order to receive full credit, you must work in a group of two to four students for at least 50 minutes in your assigned discussion section. You can also self-organize a group and meet outside of discussion section for 80 percent credit. You may not do the groupwork alone.

Problem 1.

The table below contains 12 short movie reviews that have been manually labeled as either positive ($y = 1$) or negative ($y = 0$). For every review five binary features are recorded indicating whether the corresponding word appears in the text:

review i	`great`	`boring`	`plot`	`acting`	`slow`	y
1	1	0	1	1	0	1
2	1	0	1	0	0	1
3	0	0	1	1	0	1
4	1	0	0	1	0	1
5	0	1	0	0	1	0
6	0	1	1	0	1	0
7	0	0	0	1	1	0
8	1	0	0	0	1	0
9	0	1	1	1	0	0
10	1	1	1	0	0	0
11	0	1	1	1	0	1
12	0	0	1	0	1	1

Let the feature vector $\vec{X} = (X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}, X^{(5)})$ correspond to the presence of the words `great`, `boring`, `plot`, `acting`, `slow`, respectively, and let $Y \in \{0, 1\}$ denote the sentiment class.

a) Compute the probabilities $\Pr(\{Y = 1\})$ and $\Pr(\{Y = 0\})$ from the table.

Solution: From the table, there are 12 reviews in total. The positive reviews ($y = 1$) are reviews

$$1, 2, 3, 4, 11, 12,$$

so there are 6 positives and hence 6 negatives. Therefore

$$\Pr(\{Y = 1\}) = \frac{6}{12} = \frac{1}{2}, \qquad \Pr(\{Y = 0\}) = \frac{6}{12} = \frac{1}{2}.$$

b) Consider a new review whose feature vector is

$$\vec{x} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix}^\top.$$

Using the naïve Bayes classifier, compute the class likelihoods $\Pr\left(\{Y=0\} \mid \{\vec{X}=\vec{x}\}\right)$ and $\Pr\left(\{Y=1\} \mid \{\vec{X}=\vec{x}\}\right)$. If you need to use Laplace smoothing, clearly indicate why.

Solution: The naïve Bayes classifier uses

$$\Pr\left(\{Y=y \mid \vec{X}=\vec{x}\}\right) \propto \Pr\left(\{Y=y\}\right) \prod_{j=1}^{5} \Pr\left(\{X^{(j)}=x^{(j)} \mid Y=y\}\right),$$

assuming conditional independence of the features given Y.

We first compute the necessary conditional probabilities from the table.

Among the 6 positive reviews ($Y=1$) we count:

$$\begin{aligned}
&\Pr\left(\{X^{(1)}=1 \mid Y=1\}\right)=\frac{3}{6}=\frac{1}{2}, \quad \Pr\left(\{X^{(2)}=0 \mid Y=1\}\right)=\frac{5}{6},\\
&\Pr\left(\{X^{(3)}=1 \mid Y=1\}\right)=\frac{5}{6}, \qquad \Pr\left(\{X^{(4)}=0 \mid Y=1\}\right)=\frac{2}{6}=\frac{1}{3},\\
&\Pr\left(\{X^{(5)}=1 \mid Y=1\}\right)=\frac{1}{6}.
\end{aligned}$$

Among the 6 negative reviews ($Y=0$) we count:

$$\begin{aligned}
&\Pr\left(\{X^{(1)}=1 \mid Y=0\}\right)=\frac{2}{6}=\frac{1}{3}, \quad \Pr\left(\{X^{(2)}=0 \mid Y=0\}\right)=\frac{2}{6}=\frac{1}{3},\\
&\Pr\left(\{X^{(3)}=1 \mid Y=0\}\right)=\frac{3}{6}=\frac{1}{2}, \quad \Pr\left(\{X^{(4)}=0 \mid Y=0\}\right)=\frac{4}{6}=\frac{2}{3},\\
&\Pr\left(\{X^{(5)}=1 \mid Y=0\}\right)=\frac{4}{6}=\frac{2}{3}.
\end{aligned}$$

For $y=1$ we get

$$\begin{aligned}
\Pr\left(\{Y=1,\vec{X}=\vec{x}\}\right) &= \Pr\left(\{Y=1\}\right) \prod_{j=1}^{5} \Pr\left(\{X^{(j)}=x^{(j)} \mid Y=1\}\right)\\
&= \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{3}\cdot\frac{1}{6}\\
&= \frac{25}{2592}.
\end{aligned}$$

For $y=0$ we get

$$\begin{aligned}
\Pr\left(\{Y=0,\vec{X}=\vec{x}\}\right) &= \Pr\left(\{Y=0\}\right) \prod_{j=1}^{5} \Pr\left(\{X^{(j)}=x^{(j)} \mid Y=0\}\right)\\
&= \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{2}{3}\cdot\frac{2}{3}\\
&= \frac{1}{81}.
\end{aligned}$$

The normalizing constant is

$$\Pr\left(\{\vec{X}=\vec{x}\}\right)=\Pr\left(\{Y=0,\vec{X}=\vec{x}\}\right)+\Pr\left(\{Y=1,\vec{X}=\vec{x}\}\right)=\frac{1}{81}+\frac{25}{2592}=\frac{57}{2592}.$$

Thus the posterior probabilities are

$$\Pr\left(\{Y=0 \mid \vec{X}=\vec{x}\}\right) = \frac{\frac{1}{81}}{\frac{57}{2592}} = \frac{32}{57} \approx 0.56, \qquad \Pr\left(\{Y=1 \mid \vec{X}=\vec{x}\}\right) = \frac{\frac{25}{2592}}{\frac{57}{2592}} = \frac{25}{57} \approx 0.44.$$

In this data set, every feature value (0 and 1) occurs at least once for each class, so none of the conditional probabilities above is zero. Therefore **Laplace smoothing is not needed** for this example.

c) Which class does the classifier predict for this review?

Solution: Since

$$\Pr\left(\{Y=0 \mid \vec{X}=\vec{x}\}\right) = \frac{32}{57} > \Pr\left(\{Y=1 \mid \vec{X}=\vec{x}\}\right) = \frac{25}{57},$$

the naïve Bayes classifier predicts class $Y=0$, i.e., a **negative** review.

Problem 2.

Researchers on planet Zog are building a classifier that decides whether a brand new jelly bean is Yummy ($y=1$) or Yucky ($y=0$). For each candy they record five quirky binary features:

Feature j	Meaning of $\{X^{(j)}=1\}$
1	the bean glows in the dark (`glow`)
2	the bean fizzes when bitten (`fizz`)
3	the bean feels slimy (`slimy`)
4	the bean crunches loudly (`crunch`)
5	the bean whistles when shaken (`whistle`)

The training data of 14 beans are listed below.

bean i	glow	fizz	slimy	crunch	whistle	y
1	1	1	0	1	0	1
2	1	0	0	1	0	1
3	0	1	0	0	0	1
4	1	1	0	0	0	1
5	0	0	1	0	1	0
6	0	1	1	1	0	0
7	0	0	1	0	0	0
8	1	0	1	0	1	0
9	0	1	1	0	1	0
10	1	0	1	1	1	0
11	0	1	0	1	0	1
12	1	0	0	0	0	1
13	0	0	1	1	1	0
14	1	1	1	0	1	0

Let $\vec{X} = (X^{(1)}, \ldots, X^{(5)})$ denote the five features and $Y \in \{0,1\}$ the label.

a) Compute the probabilities $\Pr(\{Y=1\})$ and $\Pr(\{Y=0\})$ from the table.

Solution: Counting labels in the table, there are 6 Yummy beans ($y=1$) and 8 Yucky beans ($y=0$)

out of 14 total. Thus

$$\Pr(\{Y=1\}) = \frac{6}{14} = \frac{3}{7}, \qquad \Pr(\{Y=0\}) = \frac{8}{14} = \frac{4}{7}.$$

b) A brand-new jelly-bean has feature vector

$$\vec{x} \;=\; \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \end{bmatrix}^\top \quad (\texttt{fizz} + \texttt{whistle} \text{ only}).$$

Estimate the probabilities $\Pr\Big(\{Y=0\} \mid \{\vec{X}=\vec{x}\}\Big)$ and $\Pr\Big(\{Y=1\} \mid \{\vec{X}=\vec{x}\}\Big)$. If you need to use Laplace smoothing, clearly indicate why.

Solution: As before, the naïve Bayes classifier uses

$$\Pr\Big(\{Y=y \mid \vec{X}=\vec{x}\}\Big) \;\propto\; \Pr(\{Y=y\}) \prod_{j=1}^{5} \Pr\Big(\{X^{(j)}=x^{(j)} \mid Y=y\}\Big).$$

Look at the raw counts from the table:

- Among the 8 YUCKY beans ($Y=0$), *all* of them have $\texttt{slimy}=1$, so

$$\Pr\Big(\{X^{(3)}=0 \mid Y=0\}\Big) = 0.$$

- Among the 6 YUMMY beans ($Y=1$), *none* have $\texttt{whistle}=1$, so

$$\Pr\Big(\{X^{(5)}=1 \mid Y=1\}\Big) = 0.$$

Our new bean has $X^{(3)}=0$ and $X^{(5)}=1$, so with unsmoothed estimates we would get

$$\Pr\Big(\{\vec{X}=\vec{x} \mid Y=0\}\Big) = 0 \quad \text{and} \quad \Pr\Big(\{\vec{X}=\vec{x} \mid Y=1\}\Big) = 0,$$

leading to zero likelihood for *both* classes. This is clearly unreasonable, so we **must use Laplace smoothing**.

For each class y and feature j, let N_y be the number of training beans with label y, and let $N_{1,y}^{(j)}$ be the number with $X^{(j)}=1$ and label y. With add-one smoothing for Bernoulli features,

$$\Pr\Big(\{X^{(j)}=1 \mid Y=y\}\Big) = \frac{N_{1,y}^{(j)}+1}{N_y+2}, \qquad \Pr\Big(\{X^{(j)}=0 \mid Y=y\}\Big) = \frac{N_y - N_{1,y}^{(j)}+1}{N_y+2}.$$

For $Y = 0$ there are $N_0 = 8$ beans, so the denominator is $8 + 2 = 10$. Counting from the table:

$$\begin{aligned}
&\text{(glow)} && N_{1,0}^{(1)} = 3, && \Rightarrow \Pr\Big(\{X^{(1)} = 0 \mid Y = 0\}\Big) = \frac{3}{5},\ \Pr\Big(\{X^{(1)} = 1 \mid Y = 0\}\Big) = \frac{2}{5};\\
&\text{(fizz)} && N_{1,0}^{(2)} = 3, && \Rightarrow \Pr\Big(\{X^{(2)} = 0 \mid Y = 0\}\Big) = \frac{3}{5},\ \Pr\Big(\{X^{(2)} = 1 \mid Y = 0\}\Big) = \frac{2}{5};\\
&\text{(slimy)} && N_{1,0}^{(3)} = 8, && \Rightarrow \Pr\Big(\{X^{(3)} = 0 \mid Y = 0\}\Big) = \frac{1}{10},\ \Pr\Big(\{X^{(3)} = 1 \mid Y = 0\}\Big) = \frac{9}{10};\\
&\text{(crunch)} && N_{1,0}^{(4)} = 3, && \Rightarrow \Pr\Big(\{X^{(4)} = 0 \mid Y = 0\}\Big) = \frac{3}{5},\ \Pr\Big(\{X^{(4)} = 1 \mid Y = 0\}\Big) = \frac{2}{5};\\
&\text{(whistle)} && N_{1,0}^{(5)} = 6, && \Rightarrow \Pr\Big(\{X^{(5)} = 0 \mid Y = 0\}\Big) = \frac{3}{10},\ \Pr\Big(\{X^{(5)} = 1 \mid Y = 0\}\Big) = \frac{7}{10}.
\end{aligned}$$

For $Y = 1$ there are $N_1 = 6$ beans, so the denominator is $6 + 2 = 8$. Counting:

$$\begin{aligned}
&\text{(glow)} && N_{1,1}^{(1)} = 4, && \Rightarrow \Pr\Big(\{X^{(1)} = 0 \mid Y = 1\}\Big) = \frac{3}{8},\ \Pr\Big(\{X^{(1)} = 1 \mid Y = 1\}\Big) = \frac{5}{8};\\
&\text{(fizz)} && N_{1,1}^{(2)} = 4, && \Rightarrow \Pr\Big(\{X^{(2)} = 0 \mid Y = 1\}\Big) = \frac{3}{8},\ \Pr\Big(\{X^{(2)} = 1 \mid Y = 1\}\Big) = \frac{5}{8};\\
&\text{(slimy)} && N_{1,1}^{(3)} = 0, && \Rightarrow \Pr\Big(\{X^{(3)} = 0 \mid Y = 1\}\Big) = \frac{7}{8},\ \Pr\Big(\{X^{(3)} = 1 \mid Y = 1\}\Big) = \frac{1}{8};\\
&\text{(crunch)} && N_{1,1}^{(4)} = 3, && \Rightarrow \Pr\Big(\{X^{(4)} = 0 \mid Y = 1\}\Big) = \frac{1}{2},\ \Pr\Big(\{X^{(4)} = 1 \mid Y = 1\}\Big) = \frac{1}{2};\\
&\text{(whistle)} && N_{1,1}^{(5)} = 0, && \Rightarrow \Pr\Big(\{X^{(5)} = 0 \mid Y = 1\}\Big) = \frac{7}{8},\ \Pr\Big(\{X^{(5)} = 1 \mid Y = 1\}\Big) = \frac{1}{8}.
\end{aligned}$$

For $\vec{x} = (0, 1, 0, 0, 1)^\top$ we obtain

$$\begin{aligned}
\Pr\Big(\{Y = 0, \vec{X} = \vec{x}\}\Big) &= \Pr\left(\{Y = 0\}\right) \Pr\Big(\{X^{(1)} = 0 \mid Y = 0\}\Big) \Pr\Big(\{X^{(2)} = 1 \mid Y = 0\}\Big) \Pr\Big(\{X^{(3)} = 0 \mid Y = 0\}\Big)\\
&\qquad \cdot \Pr\Big(\{X^{(4)} = 0 \mid Y = 0\}\Big) \Pr\Big(\{X^{(5)} = 1 \mid Y = 0\}\Big)\\
&= \frac{4}{7} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{10} \cdot \frac{3}{5} \cdot \frac{7}{10}\\
&= \frac{18}{3125} \approx 0.00576,
\end{aligned}$$

and

$$\begin{aligned}
\Pr\Big(\{Y = 1, \vec{X} = \vec{x}\}\Big) &= \Pr\left(\{Y = 1\}\right) \Pr\Big(\{X^{(1)} = 0 \mid Y = 1\}\Big) \Pr\Big(\{X^{(2)} = 1 \mid Y = 1\}\Big) \Pr\Big(\{X^{(3)} = 0 \mid Y = 1\}\Big)\\
&\qquad \cdot \Pr\Big(\{X^{(4)} = 0 \mid Y = 1\}\Big) \Pr\Big(\{X^{(5)} = 1 \mid Y = 1\}\Big)\\
&= \frac{3}{7} \cdot \frac{3}{8} \cdot \frac{5}{8} \cdot \frac{7}{8} \cdot \frac{1}{2} \cdot \frac{1}{8}\\
&= \frac{45}{8192} \approx 0.00549.
\end{aligned}$$

The posterior probabilities are then

$$\Pr\Big(\{Y = 0 \mid \vec{X} = \vec{x}\}\Big) = \frac{\Pr\Big(\{Y = 0, \vec{X} = \vec{x}\}\Big)}{\Pr\Big(\{Y = 0, \vec{X} = \vec{x}\}\Big) + \Pr\Big(\{Y = 1, \vec{X} = \vec{x}\}\Big)} \approx \frac{0.00576}{0.00576 + 0.00549} \approx 0.512,$$

and

$$\Pr\left(\{Y = 1 \mid \vec{X} = \vec{x}\}\right) \approx 0.488.$$

So after Laplace smoothing, the model slightly favors class $Y = 0$.

c) Which class does the classifier predict for this jelly bean?

Solution: Since

$$\Pr\left(\{Y = 0 \mid \vec{X} = \vec{x}\}\right) \approx 0.512 > \Pr\left(\{Y = 1 \mid \vec{X} = \vec{x}\}\right) \approx 0.488,$$

the naïve Bayes classifier predicts $Y = 0$, i.e., the bean is **Yucky**.

Problem 3. Naïve Bayes as a linear classifier

Consider a binary classification problem with label $Y \in \{0, 1\}$ and d binary features $\vec{X} = (X^{(1)}, \ldots, X^{(d)})$, where each $X^{(j)} \in \{0, 1\}$. Assume the naïve Bayes model with $\{0, 1\}$ features:

$$\Pr\left(\{Y = y\}\right) = \pi_y, \quad \Pr\left(\{X^{(j)} = 1 \mid Y = y\}\right) = \theta_{j,y}, \quad j = 1, \ldots, d, \; y \in \{0, 1\},$$

and that, conditional on Y, the features $X^{(1)}, \ldots, X^{(d)}$ are independent.

a) Show that the "log-posterior odds" can be written as

$$\log \frac{\Pr\left(\{Y = 1 \mid \vec{X} = \vec{x}\}\right)}{\Pr\left(\{Y = 0 \mid \vec{X} = \vec{x}\}\right)} \;=\; w_0 + \sum_{j=1}^{d} w_j x^{(j)},$$

for suitable constants $w_0, w_1, \ldots, w_d$ that depend only on (π_y) and $(\theta_{j,y})$. Give explicit formulas for w_0 and w_j.

Solution: By Bayes' rule and the naïve Bayes assumption,

$$\Pr\left(\{Y = y \mid \vec{X} = \vec{x}\}\right) \propto \Pr\left(\{Y = y\}\right) \prod_{j=1}^{d} \Pr\left(\{X^{(j)} = x^{(j)} \mid Y = y\}\right).$$

For Bernoulli features,

$$\Pr\left(\{X^{(j)} = x^{(j)} \mid Y = y\}\right) = \theta_{j,y}^{x^{(j)}} (1 - \theta_{j,y})^{\,1 - x^{(j)}}.$$

Thus

$$\begin{aligned}
\log \frac{\Pr\left(\{Y = 1 \mid \vec{X} = \vec{x}\}\right)}{\Pr\left(\{Y = 0 \mid \vec{X} = \vec{x}\}\right)} &= \log \frac{\pi_1}{\pi_0} + \sum_{j=1}^{d} \log \frac{\Pr\left(\{X^{(j)} = x^{(j)} \mid Y = 1\}\right)}{\Pr\left(\{X^{(j)} = x^{(j)} \mid Y = 0\}\right)} \\
&= \log \frac{\pi_1}{\pi_0} + \sum_{j=1}^{d} \log \frac{\theta_{j,1}^{x^{(j)}} (1 - \theta_{j,1})^{1 - x^{(j)}}}{\theta_{j,0}^{x^{(j)}} (1 - \theta_{j,0})^{1 - x^{(j)}}}.
\end{aligned}$$

For each j,

$$\begin{aligned}
&\log \frac{\theta_{j,1}^{x^{(j)}}(1-\theta_{j,1})^{1-x^{(j)}}}{\theta_{j,0}^{x^{(j)}}(1-\theta_{j,0})^{1-x^{(j)}}} \\
&= x^{(j)} \log \frac{\theta_{j,1}}{\theta_{j,0}} + (1-x^{(j)}) \log \frac{1-\theta_{j,1}}{1-\theta_{j,0}} \\
&= \log \frac{1-\theta_{j,1}}{1-\theta_{j,0}} + x^{(j)} \left[\log \frac{\theta_{j,1}}{\theta_{j,0}} - \log \frac{1-\theta_{j,1}}{1-\theta_{j,0}} \right] \\
&= \log \frac{1-\theta_{j,1}}{1-\theta_{j,0}} + x^{(j)} \log \frac{\theta_{j,1}(1-\theta_{j,0})}{\theta_{j,0}(1-\theta_{j,1})}.
\end{aligned}$$

Plugging back in, the log-odds is

$$\log \frac{\Pr\left(\{Y=1 \mid \vec{X}=\vec{x}\}\right)}{\Pr\left(\{Y=0 \mid \vec{X}=\vec{x}\}\right)} = \log \frac{\pi_1}{\pi_0} + \sum_{j=1}^{d} \log \frac{1-\theta_{j,1}}{1-\theta_{j,0}} + \sum_{j=1}^{d} x^{(j)} \log \frac{\theta_{j,1}(1-\theta_{j,0})}{\theta_{j,0}(1-\theta_{j,1})}.$$

This has the desired linear form

$$w_0 + \sum_{j=1}^{d} w_j x^{(j)}$$

where

$$w_0 = \log \frac{\pi_1}{\pi_0} + \sum_{j=1}^{d} \log \frac{1-\theta_{j,1}}{1-\theta_{j,0}}, \qquad w_j = \log \frac{\theta_{j,1}(1-\theta_{j,0})}{\theta_{j,0}(1-\theta_{j,1})}.$$

b) Conclude that the naïve Bayes classifier with Bernoulli features has a *linear* decision rule of the form

$$\hat{y}(\vec{x}) = \begin{cases} 1 & \text{if } w_0 + \sum_{j=1}^{d} w_j x^{(j)} \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Explain in words what this means geometrically about the decision boundary in $\{0,1\}^d$.

Solution: The decision rule predicts the class with larger posterior probability:

$$\hat{y}(\vec{x}) = \begin{cases} 1 & \text{if } \Pr\left(\{Y=1 \mid \vec{X}=\vec{x}\}\right) \geq \Pr\left(\{Y=0 \mid \vec{X}=\vec{x}\}\right), \\ 0 & \text{otherwise.} \end{cases}$$

This is equivalent to

$$\hat{y}(\vec{x}) = \begin{cases} 1 & \text{if } \log \dfrac{\Pr\left(\{Y=1 \mid \vec{X}=\vec{x}\}\right)}{\Pr\left(\{Y=0 \mid \vec{X}=\vec{x}\}\right)} \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

From part (a) we have

$$\log \frac{\Pr\left(\{Y=1 \mid \vec{X}=\vec{x}\}\right)}{\Pr\left(\{Y=0 \mid \vec{X}=\vec{x}\}\right)} = w_0 + \sum_{j=1}^{d} w_j x^{(j)},$$

so

$$\hat{y}(\vec{x}) = \begin{cases} 1 & \text{if } w_0 + \sum_{j=1}^{d} w_j x^{(j)} \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

as claimed.

If we view $\vec{x}$ as a point in $\mathbb{R}^d$ (even though its coordinates are restricted to $\{0,1\}$), the set of points satisfying

$$w_0 + \sum_{j=1}^{d} w_j x^{(j)} = 0$$

is an affine hyperplane. The classifier predicts 1 on one side of this hyperplane and 0 on the other side. Thus, in the discrete cube $\{0,1\}^d$, the naïve Bayes classifier corresponds to a *linear threshold* decision boundary: it separates the vertices of the cube into two classes using a single hyperplane.

c) Suppose for some feature j we have $\theta_{j,1} = \theta_{j,0}$. Show that $w_j = 0$ in this case and interpret this fact: what does naïve Bayes do with a feature that is *equally distributed* in both classes?

Solution: From part (a),

$$w_j = \log \frac{\theta_{j,1}(1-\theta_{j,0})}{\theta_{j,0}(1-\theta_{j,1})}.$$

If $\theta_{j,1} = \theta_{j,0}$, then

$$\theta_{j,1}(1-\theta_{j,0}) = \theta_{j,0}(1-\theta_{j,1}),$$

so the fraction inside the logarithm is 1 and hence

$$w_j = \log 1 = 0.$$

Therefore feature j does not appear in the decision rule:

$$w_0 + \sum_{k=1}^{d} w_k x^{(k)} \quad \text{does not depend on } x^{(j)}.$$

In other words, if a feature has exactly the same conditional distribution in both classes (it is equally likely to be 1 given $Y = 0$ as given $Y = 1$), then naïve Bayes assigns it zero weight and effectively *ignores* it when making predictions. Such a feature carries no discriminative information about the label in this model.