Text Classification using Naive Bayes Classifier

Tahrima Rahman

THE UNIVERSITY OF TEXAS AT DALLAS
Erik Jonsson School of Engineering and Computer Science

Feature Engineering: Case 2 (Bernoulli Features)

- If a word (feature) appears in a document (e.g. email, article, review) we assign it the value 1 (presence), otherwise we assign it value 0 (absence).
- Position of the word in the document does not matter.
- **Vocabulary size** $= n$, so number of features $= n$.
- **Example:** Suppose the vocabulary is $\{\texttt{love}, \texttt{fishing}, \texttt{music}\}$.
  - Document 1: "I love fishing."

    $$\texttt{Feature vector:} \quad \mathbf{x}^{(1)} = [1, 1, 0]$$

  - Document 2: "I love fishing. I love fishing. I love fishing..." repeated 1000 times.

    $$\texttt{Feature vector:} \quad \mathbf{x}^{(2)} = [1, 1, 0]$$

  Both documents have the same feature values.
- When is this useful?
  - Works when the **presence** of a word is as informative as its frequency. For example: presence of the word `''lottery''` may be enough to classify an email as spam.

Case 2: Parameter Estimation

- Given training data $\mathbf{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{d}$ with $\mathbf{x}^{(i)} \in \{0,1\}^n$:
- **Prior Probability of a Class y:**

$$\hat{P}(Y = y) = \frac{\texttt{\# of documents in class y}}{d}$$

- **Conditional Probability of Word X_j in a class $Y = y$:**

$$\hat{P}(X_j = 1 \mid Y = y) = \alpha_{j,y} = \frac{\texttt{\# of docs in class } y \texttt{ where word } j \texttt{ appears}\} + 1}{\texttt{total \# of documents in class y} + 2}$$

$$\hat{P}(X_j = 0 \mid Y = y) = 1 - \hat{P}(X_j = 1 \mid Y = y)$$

  - We add $+1$ (Laplace smoothing) to avoid zero probabilities.
  - Denominator uses $+2$ since $X_j \in \{0,1\}$ has two possible values.
- **Document log-likelihood:** a new document $\mathbf{x}$ has a log-likelihood

$$\log(\hat{P}(\mathbf{x} \mid Y = y)) \propto \log\left(\prod_{j=1}^{n} \alpha_{j,y}^{x_j} (1 - \alpha_{j,y})^{1-x_j}\right)$$

$$= \sum_{j=1}^{n} x_j \log(\alpha_{j,y}) + (1 - x_j)\log(1 - \alpha_{j,y})$$

where $x_j \in \{0,1\}$ indicates whether word j appears in the document. If a word appears ($x_j = 1$), its contribution is $\log(\alpha_{j,y})$. If it does not appear ($x_j = 0$), its contribution is $\log(1 - \alpha_{j,y})$.

Feature Engineering: Case 3 (Bag of Words)

- Each feature X_j represents the **number of times** word j appears in a document.
- Position of the word does not matter.
- **Vocabulary size** is same as Bernoulli NB $\rightarrow n$.
- **Compare with Case 2:**
  - In Case 2 (Bernoulli), each feature $X_j \in \{0, 1\}$ (word is present or absent).
  - In Case 3 (BoW), each feature $X_j \in \{0, 1, 2, \ldots, m\}$, where m is the maximum document length.
- **Example** vocabulary is $\{love, fishing, music\}$:
  - Document 1: "I love fishing."

$$\texttt{Feature vector:} \quad \mathbf{x}^{(1)} = [1, 1, 0]$$

  - Document 2: "I love fishing." repeated 1000 times

$$\texttt{Feature vector:} \quad \mathbf{x}^{(2)} = [1000, 1000, 0]$$

Case 3: Parameter Estimation

- Given training data $\mathbf{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{d}$, where $x^{(i)}$ is a document represented as word counts over a vocabulary of size n.
- **Prior Probability of a Class y:**

$$\hat{P}(Y = y) = \frac{\#\text{ of documents in class } y}{d}$$

- **Conditional Probability of Word X_j in a Class $Y = y$:**

$$\hat{P}(X_j \mid Y = y) = \theta_{j,y} = \frac{\#\text{ of times word } j \text{ appears in all docs of class } y + 1}{\#\text{ of total word occurrences in all docs of class } y + n}$$

  - We add $+1$ (Laplace smoothing) to avoid zero probabilities.
  - Denominator uses $+n$ since the vocabulary has n possible words.
- **Document log-likelihood:** a new document $\mathbf{x}$ has a log-likelihood

$$\log\left(\hat{P}(\mathbf{x} \mid Y = y)\right) \propto \sum_{j=1}^{n} \log(\theta_{j,y}^{d_{j,y}}) = \sum_{j=1}^{n} d_{j,y} \log(\theta_{j,y})$$

Each occurrence of a word contributes one factor of $\log(\theta_{j,y})$. If a word appears $d_{j,y}$ times, its contribution is $d_{j,y} \log(\theta_{j,y})$.

Naive Bayes: Test Document Classification

Given a test document $\mathcal{D}$:

- **Bernoulli Naive Bayes:**
  - Represent $\mathcal{D}$ as a binary vector $\mathbf{x} \in \{0,1\}^n$ (word present or absent).
  - Compute $\log(\hat{P}(Y = y|\mathbf{x}))$ for each class y:

$$\log(\hat{P}(Y = y|\mathbf{x})) = \log(\hat{P}(Y = y)) + \sum_{j=1}^{n} x_j \log(\alpha_{j,y}) + (1 - x_j)\log(1 - \alpha_{j,y})$$

  - Predict $\hat{y} = \arg\max_y \log(\hat{P}(Y = y|\mathbf{x}))$.
- **Multinomial Naive Bayes:**
  - Represent $\mathcal{D}$ as vector $\mathbf{x}$ of counts where word X_j appear d_j times in $\mathcal{D}$. (number of times word X_j appears).
  - Compute $\log(\hat{P}(Y = y|\mathbf{x}))$ for each class y:

$$\log(\hat{P}(Y = y|\mathbf{x})) = \log(\hat{P}(Y = y)) + \sum_{j=1}^{n} d_j \log\left(\theta_{j,y}\right)$$

  - Predict $\hat{y} = \arg\max_y \log(\hat{P}(Y = y|\mathbf{x}))$.

Multinomial Naive Bayes: Example (Credit: Dan Jurafsky)

► **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$
\begin{aligned}
\hat{P}(\text{Chinese}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{Chinese}|\bar{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{Tokyo}|\bar{c}) = \hat{P}(\text{Japan}|\bar{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}
$$

$$
\begin{aligned}
\hat{P}(c|d_5) &\propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003. \\
\hat{P}(\bar{c}|d_5) &\propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001.
\end{aligned}
$$

Bernoulli Naive Bayes: Example (log-scale calculations)

- Vocabulary = {Chinese, Beijing, Shanghai, Macao, Tokyo, Japan}.
- Prior: $P(Y=c) = 3/4, \quad P(Y=\bar{c}) = 1/4$
- Conditional probabilities with Laplace smoothing:

$$\hat{\alpha}_{1,c} = \tfrac{3+1}{3+2} = \tfrac{4}{5}, \quad \hat{\alpha}_{2,c} = \hat{\alpha}_{3,c} = \hat{\alpha}_{4,c} = \tfrac{2}{5}, \quad \hat{\alpha}_{5,c} = \hat{\alpha}_{6,c} = \tfrac{1}{5}$$

$$\hat{\alpha}_{1,\bar{c}} = \hat{\alpha}_{5,\bar{c}} = \hat{\alpha}_{6,\bar{c}} = \tfrac{2}{3}, \quad \hat{\alpha}_{2,\bar{c}} = \hat{\alpha}_{3,\bar{c}} = \hat{\alpha}_{4,\bar{c}} = \tfrac{1}{3}$$

- Test document: "Chinese Chinese Tokyo Japan" $\rightarrow \mathbf{x} = [1,0,0,0,1,1]$
- Document likelihood:

$$\log \hat{P}(c \mid \mathbf{x}) \propto \log(\tfrac{3}{4}) + \log(\tfrac{4}{5}) + 3\log(1 - \tfrac{2}{5}) + 2\log(\tfrac{1}{5})$$

$$\log \hat{P}(\bar{c}|\mathbf{x}) \propto \log(\tfrac{1}{4}) + 3\log(\tfrac{2}{3}) + 3\log(1 - \tfrac{1}{3})$$

- Prediction: choose class with larger posterior.

$$\log \mathrm{score}(c) \approx -5.2622 \quad \Rightarrow \quad e^{-5.2622} \approx 0.00518$$

$$\log \mathrm{score}(\bar{c}) \approx -3.8191 \quad \Rightarrow \quad e^{-3.8191} \approx 0.02195$$

Prediction, $\hat{y} = \bar{c}$ (not China)