Lecture 15

Today's topics: Finish the rest of Naïve Bayes

- Spam Filter using Naïve Bayes and multinomial Features
- Continuous Features (Guassian Naïve Bayes)

---

Text data (email) → BoW. representation (feature vector)

"An ant is an animal" → $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{pmatrix}$ $\begin{array}{l} w_1 = a \\ w_2 = an \\ w_3 = animal \\ \\ w_d = zebra \end{array}$

Bayes classifier → use information regarding $P(\underset{\sim}{y} \mid \underset{\sim}{x})$ spam/ham, email

Naïve Bayes → view $\boxed{P(Y \mid \vec{x}) = \dfrac{P(\vec{x} \mid y)\, P(y)}{P(x)}}$

$\Rightarrow P(Y \mid x) \propto \underset{\sim}{P(x \mid y)}\; \underset{\sim}{P(y)}$

By Naïve Bayes assumption,

$$P(X \mid Y) = \prod_{\alpha=1}^{d} P(\underset{\sim}{x_\alpha} \mid \underset{\sim}{Y})$$

$Y$: spam/ham

$x_\alpha$: $w_\alpha$ appears $x_\alpha$ times

In binary classification

$$P(Y=y) \approx \frac{\sum_{i=1}^{n} I(y_i = y)}{n}$$

Training: - Estimate $P(Y)$

- Estimate $P(X_\alpha \mid Y)$ for each feature $\alpha$

"An ant is an animal"

"An animal is an ant"

$$\vec{x} = \begin{array}{c|l} \begin{pmatrix} 0 \\ 2 \\ 1 \\ 1 \\ \vdots \\ 0 \end{pmatrix} & \begin{array}{l} w_1 = a \\ w_2 = an \\ w_3 = animal \\ w_4 = ant \\ \vdots \\ w_d = zebra \end{array} \end{array}$$

$$P(X_2 = 2 \mid y = SPAM) = \binom{5}{2} \cdot P(w_2 \mid y = SPAM)^2$$

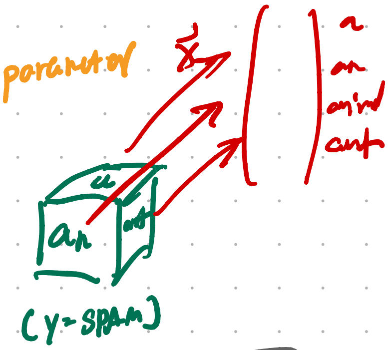
In general, $P(X_\alpha = c \mid y = SPAM) = \binom{m}{c} \cdot P(w_\alpha \mid y = SPAM)^c$

, where $P(w_\alpha \mid Y = SPAM)$ is the probability of selecting the word $w_\alpha$ and $m$ is the number of words in the email

$m = \sum_{\alpha=1} x_\alpha$

## Probabilistic models for spam Filter:

model parameter




- Assume $P_\theta(w_\alpha \mid y = spam) = [\theta_{SPAM}]_\alpha$

$P_\theta(w_\alpha \mid y = ham) = [\theta_{HAM}]_\alpha$

Training

prior

Estimate $[\theta_{SPAM}]_\alpha = \dfrac{\sum_{i=1}^{n} I(y_i = spam)\, x_{i\alpha} + l}{\sum_{i=1}^{n} I(y_i = spam)\left(\sum_{\alpha=1}^{d} m_i\right) + l \cdot d}$

Likewise, we do the same for $[\theta_{HAM}]_\alpha$

$$\Rightarrow P_\theta(x_\alpha = c \mid y = spam) = \binom{m}{c}\left([\theta_{SPAM}]_\alpha\right)^c$$

Testing $\rightarrow$ find $y$ s.t. $P_\theta(\vec{x} \mid y)\, P_\theta(y)$ is maximized

$$P_\theta(\vec{x} \mid y = spam) = \prod_{\alpha=1}^{d} P_\theta(x_\alpha \mid y = SPAM)$$

$$= \binom{m}{x_1}\left([\theta_{spam}]_1\right)^{x_1} \times \binom{m-x_1}{x_2}\left([\theta_{spam}]_2\right)^{x_2} \times \cdots \times \binom{m-(x_1+\dots+x_{d-1})}{x_d}\left([\theta_{spam}]_d\right)^{x_d}$$

$$= \prod_{\alpha=1}^{d}\left([\theta_{spam}]_\alpha\right)^{x_\alpha} \times \binom{m}{x_1} \times \binom{m-x_1}{x_2} \times \cdots \times \binom{m-(x_1+\dots+x_{d-1})}{x_d}$$

$$= \prod_{\alpha=1}^{d}\left([\theta_{spam}]_\alpha\right)^{x_\alpha} \times \frac{m!}{x_1!\,\cancel{(m-x_1)!}} \times \frac{\cancel{(m-x_1)}!}{x_2!\,\cancel{(M-(x_1+x_2))}!} \times \cdots \times \frac{\cancel{(m-(x_1+\dots+x_{d-1}))}!}{x_d!\,\underbrace{\cancel{(m-(x_1+\dots+x_d))!}}_{=0}}$$

$$= \prod_{\alpha=1}^{d}\left([\theta_{spam}]_\alpha\right)^{x_\alpha} \times \frac{m!}{x_1!\,x_2!\cdots x_d!}$$

Trick to avoid computing this term

$$\frac{P(\vec{x} \mid y = SPAM)}{P(\vec{x} \mid y = ham)} = \frac{\prod_{\alpha=1}^{d}\left([\theta_{SPAM}]_\alpha\right)^{x_\alpha} \times \cancel{\frac{m!}{x_1!\,x_2!\cdots x_d!}}}{\prod_{\alpha=1}^{d}\left([\theta_{HAM}]_\alpha\right)^{x_\alpha} \times \cancel{\frac{m!}{x_1!\,x_2!\cdots x_d!}}}$$

if the ratio $\geq 1$, answer "spam"

otherwise, answer "ham"

# LET'S MAKE OUR SPAM FILTER.

## Multinomial Features:

a count for feature $\alpha$ → $x_\alpha \in \{0, 1, 2, \dots, m\}$ and $\sum_{\alpha=1}^{d} x_\alpha = m$ ← total count across d features


⇒ $\underbrace{b_1, b_1, b_2, b_2, b_2, b_3}$ ⇒ $x = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$


---

## Continuous Features (Gaussian Naive Bayes)

$x_\alpha \in \mathbb{R}$ (e.g. heights, weights)

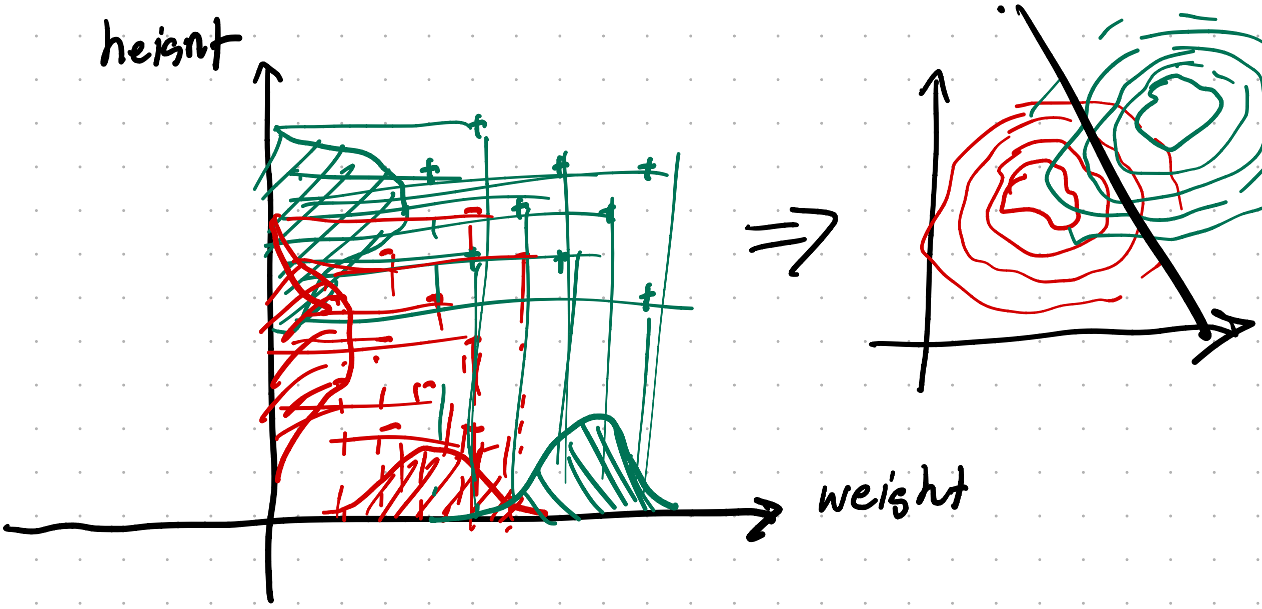
- Probabilistic model

  - Assume $P_\theta(x_\alpha \mid y) = \underline{\mathcal{N}([\mu_y]_\alpha, [\sigma_y]_\alpha)}$

    Gaussian w/ mean $[\mu_y]_\alpha$ and s.d. $[\sigma_y]_\alpha$ as parameters

  - We already know how to estimate $P_\theta(x_\alpha \mid y)$

# Gaussian Naive Bayes




## Summary of Naive Bayes

- Naive Bayes = Bayes classifier + Naive Bayes assumption
- The assumption is "all feature values are independent."
- The cons. is that we might the data that breaks the assumption. (e.g. image data)
- In most cases, Naive bayes give linear decision boundary
  (Naive Bayes is a linear classifier)