10-301/601: Introduction to Machine Learning Lecture 17.5 - Naïve Bayes Predictions

Henry Chai

11/1/22

Bernoulli Naïve Bayes

- Binary label
  - $Y \sim \text{Bernoulli}(\pi)$
  - $\hat{\pi} = {}^{N_{Y=1}}/_{N}$
    - N = # of data points
    - $N_{Y=1}$ = # of data points with label 1
- Binary features
  - $X_d | Y = y \sim \text{Bernoulli}(\theta_{d,y})$
  - $\hat{\theta}_{d,y} = {}^{N_{Y=y,\, X_d=1}}/_{N_{Y=y}}$
    - $N_{Y=y}$ = # of data points with label y
    - $N_{Y=y,\, X_d=1}$ = # of data points with label y and feature $X_d = 1$

What if some Word-Label pair never appears in our training data?

Bernoulli Naïve Bayes: Making Predictions

- Given a test data point $\boldsymbol{x}' = [x_1', \dots, x_D']^T$

$$P(Y=1 \mid x') \propto P(x' \mid Y=1)\, P(Y=1)$$

$$= \left(\prod_{d=1}^{D} p(x_d' \mid Y=1)\right) P(Y=1)$$

$$\bigstar \quad = \left(\prod_{d=1}^{D} \hat{\theta}_{d,1}^{\,x_d'} (1-\hat{\theta}_{d,1})^{1-x_d'}\right) \hat{\pi} \quad := P_1$$

$$P(Y=0 \mid x') \propto \left(\prod_{d=1}^{D} \hat{\theta}_{d,0}^{\,x_d'} (1-\hat{\theta}_{d,0})^{1-x_d'}\right)(1-\hat{\pi}) := P_0$$

$$\hat{Y} = \begin{cases} 1 & \text{if } P_1 \geq P_0 \\ 0 & \text{otherwise} \end{cases}$$

What if some Word-Label pair never appears in our training data?

x_1 ("hat")	x_2 ("cat")	x_3 ("dog")	x_4 ("fish")	x_5 ("mom")	x_6 ("dad")	y (Dr. Seuss)
1	1	0	0	0	0	1
0	0	1	0	0	0	0
0	0	0	1	0	0	1
0	0	0	0	1	0	0

The Cat in the Hat gets a Dog (by ???)

- If some $\hat{\theta}_{d,y} = 0$ and that word appears in our test data $\boldsymbol{x}'$, then $P(Y = y|\boldsymbol{x}') = 0$ even if all the other features in $\boldsymbol{x}'$ point to the label being y!
- The model has been overfit to the training data
- We can address this with a prior over the parameters!

Setting the Parameters via MAP

- Binary label
  - $Y \sim \text{Bernoulli}(\pi)$
  - $\hat{\pi} = {N_{Y=1}}/{N}$
    - N = # of data points
    - $N_{Y=1}$ = # of data points with label 1
- Binary features
  - $X_d | Y = y \sim \text{Bernoulli}(\theta_{d,y})$ and $\underline{\theta_{d,y} \sim \text{Beta}(\alpha, \beta)}$
  - $\hat{\theta}_{d,y} = {N_{Y=y,\, X_d=1} + (\alpha - 1)}/{N_{Y=y} + (\alpha - 1) + (\beta - 1)}$
    - $N_{Y=y}$ = # of data points with label y
    - $N_{Y=y,\, X_d=1}$ = # of data points with label y and feature $X_d = 1$
  - Common choice: $\alpha = 2, \beta = 2$