Generative Modeling and Naive Bayes:

DS 4400
Instructor: Ehsan Elhamifar

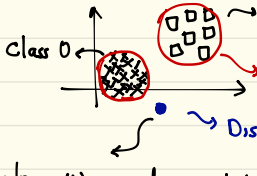
Given input features/attributes $\{x^1, \ldots, x^N\}$ and discrete outputs/labels $\{y^1, \ldots, y^N\}$, find a model to predict y given x

$x^i = \begin{pmatrix} x^i_1 \\ x^i_2 \end{pmatrix}$ → feature/attribute/covariate

$y^i \in \{0, 1, \ldots, L-1\}$

We discussed two ways to address the problem

Classification
- Discriminative Modeling: model $p(y|x)$ — How labels are influenced by features!
- Generative Modeling: model $p(y)$, $p(x|y)$
  - $p(y)$: Class prior
  - $p(x|y)$: class conditional → How features are influenced by the labels



(1) generative: which class prob is more likely to have generated the test sample

(2) discriminative: x is on which side of the decision boundary

Generative Modeling Specify $p(y)$, $p(x|y)$ and using Bayes rule

$$\underset{j \in \{0, \ldots, L-1\}}{\operatorname{argmax}} \; p(y=j \mid x) = \underset{j \in \{0, \ldots, L-1\}}{\operatorname{argmax}} \; p(y=j)\, p(x \mid y=j)$$

* We need to learn parameters of generative model via maximum likelihood, given a training dataset $D = \{(x^1, y^1), \ldots, (x^N, y^N)\}$

* what are parameters to learn ? $p(y=j) \triangleq \theta_j^y$

$p(x=\bar{x} \mid y=j) \triangleq \theta_{\bar{x},j}^{x|y}$

$\longrightarrow \theta \triangleq \{\theta_j^y, \theta_{\bar{x}|j}^{x|y} ; \forall j, \forall \bar{x}\}$

Ex) Spam detection $x = \begin{pmatrix} \text{'Free'} \\ \text{'CMS'} \\ \text{'Call now'} \\ \text{'Nigerian'} \\ \vdots \end{pmatrix}$ $y \in \{\text{'spam'}, \text{'non-spam'}\}$

Need to specify
$\theta_0^y = p(y=0) = p(\text{'spam' class})$
$\theta_1^y = p(y=1) = p(\text{'not spam' class})$
$\theta_{\bar{x}|0}^{x|y} = p(x=\bar{x} \mid y=0) = p(x=\bar{x} \mid \text{'spam'})$
$\theta_{\bar{x}|1}^{x|y} = p(x=\bar{x} \mid y=1) = p(x=\bar{x} \mid \text{'not spam'})$

* MLE $L(\theta) = P_\theta(D) = P_\theta(x^1,y^1, \dots, x^N,y^N) \overset{iid}{=} \prod_{i=1}^N P_\theta(x^i,y^i) = \prod_{i=1}^N P_\theta(y^i) \prod_{i=1}^N P_\theta(x^i \mid y^i)$

$= \prod_{i=1}^N \prod_{j=0}^{L-1} \underbrace{P_\theta(y^i=j)}_{=\theta_j^y}{}^{1(y^i=j)} \times \prod_{i=1}^N \prod_{j=0}^{L-1} \prod_{\bar{x}} \underbrace{P_\theta(x^i=\bar{x} \mid y^i=j)}_{\theta_{\bar{x}|j}^{x|y}}{}^{1(x^i=\bar{x}, y^i=j)}$

$\rightarrow \ell(\theta) = \log P_\theta(D) = \sum_{i=1}^N \sum_{j=0}^{L-1} 1(y^i=j) \log \theta_j^y + \sum_{i=1}^N \sum_{j=0}^{L-1} \sum_{\bar{x}} 1(x^i=\bar{x}, y^i=j) \log \theta_{\bar{x}|j}^{x|y}$

$= \sum_{j=0}^{L-1} \left(\sum_{i=1}^N 1(y^i=j)\right) \log \theta_j^y + \sum_{j=0}^{L-1} \sum_{\bar{x}} \left(\sum_{i=1}^N 1(x^i=\bar{x}, y^i=j)\right) \log \theta_{\bar{x}|j}^{x|y}$

Recall: $\sum_{j=1}^p \alpha_j \log \theta_j$ if $\theta_j \geq 0$, $\sum_{j=1}^p \theta_j = 1$ $\longrightarrow$ $\hat{\theta}_j = \frac{\alpha_j}{\sum_{\ell=1}^p \alpha_\ell} \Rightarrow \hat{\theta}_j^y = \frac{\sum_{i=1}^N 1(y^i=j)}{\sum_{j=0}^{L-1} \sum_{i=1}^N 1(y^i=j)} = \frac{\#\text{Samples in } j}{N}$

ex) 1520 spam emails, 1933 non-spam ema

$\theta_0^y = \frac{1520}{1520+1933}$, $\theta_1^y = \frac{1933}{1520+1933}$

Similarly, $\hat{\theta}_{\bar{x}|j}^{x|y} = \frac{\sum_{i=1}^N 1(x^i=\bar{x}, y^i=j)}{\sum_{\bar{x}} \sum_{i=1}^N 1(x^i=\bar{x}, y^i=j)} = \frac{\#\text{ Samples in class } j \text{ whose feature is } \bar{x}}{\#\text{ Samples in class } j}$

* How about specifying $\theta^{x|y}_{\bar{x}|j} \triangleq p(x=\bar{x}|y=j)$?

→ Similar to before, for each possible value of $x = \bar{x}$, we need to count the fraction of training samples in class j whose feature is $\bar{x}$!

Ex) $x = \begin{pmatrix} \text{'Free'} \\ \text{'CAPS'} \\ \text{'RBJ'} \end{pmatrix}$ for $\bar{x} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ count the fraction of emails in

class 'spam' where $x' = \bar{x}$ → $\theta^{x|y}_{\bar{x},0}$

class 'non spam' " " → $\theta^{x|y}_{\bar{x},1}$

* Challenge if x's have K features $\begin{pmatrix} x'_1 \\ x'_K \end{pmatrix}$ where each feature takes

m discrete values → we have m^K possible $\bar{x}$'s !!

↳ Large !

(in the case of spam detection example, 2^K)

We need to have a very large training set with multiple examples per each of m^K cases, to get a valid estimation of $\theta^{x|y}_{\bar{x}|j}$!!!

$L \times m^K$ pairs Impossible ! In the case of limited training we do not have any example for which $x = \bar{x}$ for many $\bar{x}$'s !

ex) $\bar{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 0 \end{pmatrix}$ ↝ have not seen this before, so estimate $\theta^{x|y}_{\bar{x}|0} = \theta^{x|y}_{\bar{x}|1} = 0$

but makes sense to have $\theta^{x|y}_{\bar{x}|0} \simeq 1$, $\theta^{x|y}_{\bar{x}|1} \simeq 0$!!

$$\underset{j=0,\dots,L-1}{\text{argmax}}\ p(y=j|x^t) = \underset{j=0,1,\dots,L-1}{\text{argmax}}\ p(x^t|y=j)\,p(y=j) = \underset{j=0,\dots,L-1}{\text{argmax}}\ \underbrace{\theta^{x|y}_{x^t|j}}_{\text{zero at training}} \times \theta^{y}_{j} = 0 \quad , \forall j$$

Drawbacks → exponential number of parameters m^K in the number of features

→ many unseen examples among m^K → poor estimate

Fix "Naive Bayes" Assumption

features are independent from each other given a class

$$p(x' \mid y=j) = p(x'_1 \mid y=j) \times \cdots \times p(x'_K \mid y=j)$$
$$= \prod_{k=1}^{K} p(x'_k \mid y=j) \quad \rightsquigarrow \text{Conditional Independence}$$

Ex) In the case of spam detection, if we know an email is a spam, the chance of seeing 'Free' does not change if we see words in 'CAPS'

Conditional Independence
$$p(\text{'Free'}=1, \text{'CAPS'}=1 \mid \text{Spam}) = p(\text{'Free'}=1 \mid \text{spam})\, p(\text{'CAPS'}=1 \mid \text{spam})$$

However, this independence does not hold if we do not condition

$$p(\text{'Free'}=1, \text{'RUN'}=1) \neq p(\text{'Free'}=1)\, p(\text{'CAPS'}=1)$$

if we see 'free', that increases the chance of email being 'spam' which consequently increases the chance of seeing 'CAPS'

So, we need to specify $p(x_k = \bar{x}_k \mid y=j) \triangleq \theta^{x_k \mid y}_{\bar{x}_k, j}$ for all $j = 0, \ldots, L-1$ and all x_k values. If x_k takes m values $\longrightarrow L \times m \times K$ parameters

$\Rightarrow$ significant decrease from $L \times m^K$ to $L \times m \times K$

Remember that for MLE:

$$P_\theta(D) = \prod_{i=1}^{N} P_\theta(y^i) \underbrace{\prod_{i=1}^{N} P_\theta(x^i|y^i)}$$

$$= \prod_{i=1}^{N}\prod_{l=1}^{K} P_\theta(x_l^i|y^i) = \prod_{i=1}^{N}\prod_{l=1}^{K}\prod_{j=0}^{L-1}\prod_{\bar{x}_l} P(x_l^i=\bar{x}_l, y^i=j)^{1(x_l^i=\bar{x}_l, y^i=j)}$$

$$\triangleq \theta_{\bar{x}_l|j}^{x_l|y}$$

$$\rightarrow \ell(\theta) = \dots + \sum_{l=1}^{K}\sum_{j=0}^{L-1}\sum_{\bar{x}_l}\left(\sum_{i=1}^{N} 1(x_l^i=\bar{x}_l, y^i=j)\right)\log \theta_{\bar{x}_l|j}^{x_l|y}$$

$$\rightarrow \hat{\theta}_{\bar{x}_l|j}^{x_l|y} = \frac{\sum_{i=1}^{N} 1(x_l^i=\bar{x}_l, y^i=j)}{\sum_{\bar{x}_l}\sum_{i=1}^{N} 1(x_l^i=\bar{x}_l, y^i=j)} = \frac{\#\text{ samples in class } j \text{ whose } l\text{-th feature is } \bar{x}_l}{\#\text{ samples in class } j}$$

Ex) $\hat{\theta}_{111}^{x_1|y} = \dfrac{\#\text{ emails in class spam that contain word 'Free'}}{\#\text{ emails in class spam}}$

Ex) Assume in none of training examples, the word 'call now' appears, but words such as 'Free', 'CAPS',... appear $\Rightarrow \hat{\theta}_{\bar{x}_3|j}^{x_3|y} = 0, \ \forall \bar{x}_3 \in \{0,1\}, j \in \{0,1\}$

Assume in test email, we have word 'call now' as well as 'Free', 'CAPS',..

$$P(y=j|x^t) \propto P(x^t|y=j)P(y=j) = P(x_1^t|y=j)\cdots \underbrace{P(x_3^t|y=j)}_{\text{since } P(x_3^t|y)=0} P(y=j) = 0$$

To overcome this zero prob effect, we assume all $\theta_{\bar{x}_l|j}^{x_l|y} \neq 0$ by defining

$$\hat{\theta}_{\bar{x}_l|j}^{x_l|y} = \frac{\sum_{i=1}^{N} 1(x_l^i=\bar{x}_l, y=j) + t}{\sum_{\bar{x}_l}\sum_{i=1}^{N} 1(x_l^i=\bar{x}_l, y=j) + mt}$$

when x_l takes m values.

↖ Laplace smoothing

→ t is a hyperparam which can be tuned using cross validation

$10^{-3}, 10^{-6}$

In the MLE framework,

$$\theta^{x_k|y}_{\bar{x},j} = \frac{\sum_{i=1}^{N} 1(x_k^i = \bar{x}, y^i = j)}{\sum_{i=1}^{N}\sum_{l=1}^{m} 1(x_{i_k}^i = l, y^i = j)}$$

↳ fraction of emails in class j, whose k-th feature is $\bar{x}$!

* Using Naive Bayes, we reduced the # parameters from exponential to linear in the # features. However, it still can be the case that we do not observe any training data for some $\theta^{x_k|y}_{\bar{x}|j}$

Ex) Bag of Words model for Documents

$$\text{Doc } i = \begin{pmatrix} 105 \to a \\ 69 \to \text{ea} \\ 28 \to \text{rf} \\ 23 \to \text{car} \\ 0 \to \text{truck} \end{pmatrix}$$

each feature can take length of document # of values $\{1, 2, \ldots, |\text{Doc}_i|\}$

$\longrightarrow |\text{Doc}|^K$ possible feature combinations / # params

$\xrightarrow{\text{NB}} K\,|\text{Doc}|$ possible parameters

Car vs politics classification of documents: assume in non of training samples in class car, there is word 'truck' $\longrightarrow p(\text{'truck'} \neq 0 \mid y = \text{'car'}) = 0$

Now, in your test if word 'truck' appear in a document of class 'car', then $p(x \mid \text{'car'}) = \prod_k p(x_k \mid \text{'car'}) = 0$!!! → You do not choose class 'car' for true class!

Solution: Laplace smoothing in the past — assume we have seen each of the words/features t times

$$\theta^{x_k|y}_{\bar{x}|j} = \frac{\sum_{i=1}^{N} 1(x_k^i = \bar{x}, y^i = j) + t}{\sum_{i=1}^{N}\sum_{l=1}^{m} 1(x_{i_k}^i = l, y^i = j) + mt}$$

Example: Spam detection with $x = \begin{pmatrix} \text{'Free'} \\ \text{'CAPS'} \\ \text{'Call now'} \end{pmatrix}$ spam: $y=1$, non-spam: $y=0$

$$D = \left\{ \left(\begin{pmatrix}1\\1\\1\end{pmatrix}, 1\right), \left(\begin{pmatrix}0\\1\\1\end{pmatrix}, 1\right), \left(\begin{pmatrix}1\\0\\1\end{pmatrix}, 1\right), \left(\begin{pmatrix}0\\1\\0\end{pmatrix}, 1\right), \left(\begin{pmatrix}1\\1\\0\end{pmatrix}, 0\right), \left(\begin{pmatrix}1\\0\\0\end{pmatrix}, 0\right), \left(\begin{pmatrix}0\\1\\0\end{pmatrix}, 0\right) \right\}$$

$$\theta^y_0 = \frac{3}{7}, \quad \theta^y_1 = \frac{4}{7}$$

$$\theta^{x_1|y}_{01} = \frac{2}{3}, \quad \theta^{x_1|y}_{11} = \frac{1}{3}$$
$$\theta^{x_2|y}_{01} = \frac{3}{3}, \quad \theta^{x_2|y}_{11} = 0$$
$$\theta^{x_3|y}_{01} = \frac{2}{3}, \quad \theta^{x_3|y}_{11} = \frac{1}{3}$$

$$\theta^{x_1|y}_{02} = \frac{2}{4}, \quad \theta^{x_1|y}_{12} = \frac{2}{4}$$
$$\theta^{x_2|y}_{02} = 1, \quad \theta^{x_2|y}_{12} = 0$$
$$\theta^{x_3|y}_{02} = \frac{3}{4}, \quad \theta^{x_3|y}_{12} = \frac{1}{4}$$

$x^n = \begin{pmatrix}1\\1\\1\end{pmatrix} \Rightarrow \underset{j=0,1}{\text{argmax}}\, p(y=j|x^n) = \underset{j=0,1}{\text{argmax}}\, p(x^n|y=j)\,p(y=j) = \underset{j=0,1}{\text{argmax}}\, p(x^n_1|y=j)\,p(x^n_2|y=j)\,p(x^n_3|y=j)\,p(y=j)$

$$\underset{j=0,1}{\text{argmax}}\ \theta^{x_1|y}_{1j}\ \theta^{x_2|y}_{1j}\ \theta^{x_3|y}_{1j}\ \theta^{y}_{j}$$

$j=0 \longrightarrow \frac{1}{3} \times \frac{\varepsilon}{0} \times \frac{1}{3} \times \frac{3}{7}$ with $0 \Rightarrow$ cannot distinguish

$j=1 \longrightarrow \frac{2}{4} \times \frac{\varepsilon}{0} \times \frac{1}{4} \times \frac{4}{7}$ with $\varepsilon \quad \frac{1}{21} < \frac{1}{14} \Longrightarrow j^* = 1$

$x^n = \begin{pmatrix}1\\0\\1\end{pmatrix} \Rightarrow \underset{j=0,1}{\text{argmax}}\, p(y=j|x^n) = \underset{j=0,1}{\text{argmax}}\, p(x^n|y=j)\,p(y=j) = \underset{j=0,1}{\text{argmax}}\, p(x^n_1=1|y=j)\,p(x^n_2=0|y=j)\,p(x^n_3=1|y=j)\,p(y=j)$

$$= \underset{j=0,1}{\text{argmax}}\ \theta^{x_1|y}_{1j}\ \theta^{x_2|y}_{0j}\ \theta^{x_3|y}_{1j}\ \theta^{y}_{j}$$

$j=0 \longrightarrow \frac{1}{3} \times 1 \times \frac{1}{3} \times \frac{3}{7} = \frac{1}{21}$

$j=1 \longrightarrow \frac{2}{4} \times 1 \times \frac{1}{4} \times \frac{4}{7} = \frac{1}{14}$ $\wedge$ $\longrightarrow j^* = 1$ ✓

Gaussian Naive Bayes

The previous examples, each feature takes a discrete set of values binary or m values
what if features are continuous $x = \begin{pmatrix} \text{time} \\ \text{location} \end{pmatrix}$? ex in weather classification

Model $p(x_k \mid y=j)$ using a continuous distribution

$$x_k \mid y=j \sim \mathcal{N}(\mu_{kj}, \sigma^2_{kj})$$

$$\rightarrow \hat{\mu}_{kj} = \frac{\sum_{i=1}^{N} x'_k \, 1(y'=j)}{\sum_{i=1}^{N} 1(y'=j)}$$

$$\rightarrow \sigma^2_{kj} = \frac{\sum_{i=1}^{N} (x'_k - \hat{\mu}_{kj})^2 \, 1(y'=j)}{\sum_{i=1}^{N} 1(y'=j)}$$