Probabilistic Classifiers

Probabilistic classifiers

Classifier outputs a probability distribution over the class labels: $X \rightarrow P(y|x)$

$\underline{x} : x_1, x_2 \dots x_d \longrightarrow$ classifier $\longrightarrow \hat{P}(y|x)$ $\quad \hat{y}$

$y \in \{1, 2, \dots k\}$

predicted label

probability distribution over class labels.

$x_j \in$ Real or Categorical

$$\hat{y} = \arg\max_{y} \hat{P}(y|x)$$

Eg $\hat{P}(y|x) = \begin{array}{ccc} 1 & 2 & 3 \\ \hline 0.2 & 0.7 & 0.1 \end{array}$

$\underline{k = 3}$

$\hat{y} = 2$

3

Types of Classifiers

- Generative: learns to generate the examples
  - Learn P(x|y) and P(y) from the training data and then apply Bayes rule to find P(y|x)

$$P(y|x) = \frac{P(y)P(x|y)}{\underbrace{\sum_{y'} P(y')P(x|y')}_{= P(x)}}$$

Eg $k=2$

$$P(y=1|x) = \frac{P(y=1)\,P(x|y=1)}{P(y=1)P(x|y=1) + P(y=2)P(x|y=2)}$$

Training time: Estimate P(y) and P(x|y) for each y = 1 -- K

Types of Classifiers

- Conditional Classifiers: Model conditional distribution $P(y|x)$ directly.
  - Example: logistic regression classifier
  - Neural Networks.

Generative classifiers

- Modeling P(y): Easy, k-way multinomial distribution for k-classes.
- Modeling P(x|y): Challenges, high-dimensional datasets
- Spaces required:

$x \in$ d-dimensional data $\quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$ eg: x_j = if pixel j in image is 1 or 0.

d could be large $\approx 128 \times 128 \equiv 2^{14}$ ≈ 16000

$P\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \middle| y \right)$ If we represent this full joint distribution, then to estimate # of parameters would be 2^d

for each 'y'.

$O(\text{parameters})$

$2^{2^{14}}$

Naive Bayes Classifier: A Generative Classifer

Each attribute x_j is conditionally independent given the class label

Formula: $P(\bar{x}|y) = P([x_1, x_2 \ldots x_d]|y) = \prod_{j=1}^{d} \boxed{P(x_j|y)}$

Example: conditioned on the digit (class label). whether pixel x_j is "1" or "0" is independent of whether any other pixel is "1" or "0"

Training a Naïve Bayesian Classifier

- Given training data, apply maximum likelihood principle to estimate parameters
- Estimating P(y): A multinomial distribution.
- Estimating $P(x_j|y)$

is also multinomial

  - If j-th attribute is categorical: P(xj|y) is estimated as the relative freq of samples having value di as j-th attribute in class y
  - If j-th attribute is continuous: P(xj|y) is estimated through a continuous density function: eg. Gaussian density function
- Computationally easy in both cases

$$P(x_j|p) \sim \mathcal{N}(x_j; \mu_{jp}; \sigma_{jp}^2)$$

$$\mu_{jp} = \sum_{i=1 \text{ if } y_i = p}^{N} (x_j^i) \Big/ \sum_{i=1}^{N} 1 \text{ if } (y_i = p) \; ; \; \sigma_{jp}$$

Play-tennis example: estimating $P(x_i|C)$

x_1 x_2 x_3 x_j $y \in \{P, N\}$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

$p(y)$

P(p) = 9/14
P(n) = 5/14

outlook $P(x_1 \mid y=p)$	$P(x_1 \mid y=n)$
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature $P(x_2 \mid p)$	$P(x_2 \mid n)$
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 2/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

Naive Bayesian Classifier (II)

- Given a training set, we can compute the probabilities

Outlook	P	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	P	N
high	3/9	4/5
normal	6/9	1/5
Windy		
true	3/9	3/5
false	6/9	2/5

'0' probabilities are a problem.
In practice we smooth estimates
eg: Lidstone or Laplace smoothing

Play-tennis example: classifying X

- An unseen sample X = <rain, hot, high, false>
- $P(X|p) \cdot P(p) =$

  $P(p|x) = \frac{P(x|p)\,P(p)}{P(x)}$

  $P(rain|p) \cdot P(hot|p) \cdot P(high|p) \cdot P(false|p) \cdot P(p)$
  - $= 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$
- $P(X|n) \cdot P(n)$

  $P(n|x) = \frac{P(x|n)\,P(n)}{P(x)}$

  $= P(rain|n) \cdot P(hot|n) \cdot P(high|n) \cdot P(false|n) \cdot P(n)$
  - $= 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286$
- Sample X is classified in class n (don't play)

$P(n|x) = 0.018286 / (0.018286 + 0.010582) \approx 0.6$

Summary of Naïve Bayes Classifier

- Creates a distribution over class label given x P(y|x) by applying Bayes rule.
    - Requires estimating P(x|y) for each class y and P(y)
- Estimates P(x|y) by assuming that each attributes of x are conditionally independent given the class label
    - Very easy computationally.
- Many applications in spite of simplistic assumption: e.g. classifying emails as spam Vs non-spam
- Limitations of generative method: estimating P(x|y) is hard since x could be high-dimensional. Useful when P(x|y) is already available, e.g. in speech recognition use of HMMs for word recognition

Demo

https://colab.research.google.com/drive/1_9j1CAvkPHq18zhZ3tBj4r3I7-LZKZ3c?usp=sharing

Conditional models

Conditional Probabilistic Approach:

- We will model the conditional distribution: P(y | x): Instead of a single value, we predict a distribution over values to reflect uncertainty.
- Example: regression models.
  - $P(y|x) \sim N(y; \mu_x, \sigma)$
  - $\mu_x = w^T.x + b = w_1 x_1 + \cdots + w_d x_d + b$
    - Mean prediction is a linear function of x.
  - $\sigma = independent\ of\ x.$
- 1-d diagram

Estimating parameters using MLE

Classification (Linear) --- Logistic Classifier

- y is binary, $x \in R^d$
- Conditional Probabilistic Approach:
    - We will model the conditional distribution: P(y | x)
    - $P(y|x) \sim Bernoulli(y; \theta_x)$
- How to obtain Bernoulli parameter (has to be between 0 and 1) from x?
    - Compute a linear function x: $g(x) = w.x + b, \;\; g(x) \in [-\infty, \infty]$
    - Use a sigmoid function to squash g(x) between 0 and 1.

Extending Logistic Regression to Multi-class

- A class label y can take one ~~of~~ k possible discrete values.
- Pr(y|x) ~ Multinomial distribution with k parameters each of which is a function of x
- $\Pr(y|x) \sim Mult(y; \{\theta_1(x), \dots, \theta_k(x)\})$; $\theta_r(x) \geq 0 \mid \sum_{r=1}^{k} \theta_r(x) = 1$

$(w^1, b^1), \dots (w^k, b^k)$ k sets of parameters. $\vec{\theta}(x)$ is a simplex.

$$\theta_r(x) = \frac{e^{w^r \cdot x + b^r}}{\sum_{r'=1}^{k} e^{w^{r'} \cdot x + b^{r'}}}$$

Softmax

Estimating parameters