Introduction to Machine Learning

Bayesian Classification

Varun Chandola

Computer Science & Engineering
State University of New York at Buffalo
Buffalo, NY, USA
chandola@buffalo.edu

University at Buffalo
Department of Computer Science and Engineering
School of Engeering and Applied Sciences

Outline

Learning Probabilistic Classifiers

Training data, $D = [\langle \mathbf{x_i}, y_i \rangle]_{i=1}^{D}$

1. {circular,large,light,smooth,thick}, malignant
2. {circular,large,light,irregular,thick}, malignant
3. {oval,large,dark,smooth,thin}, benign
4. {oval,large,light,irregular,thick}, malignant
5. {circular,small,light,smooth,thick}, benign

- **Testing**: Predict y^* for $\mathbf{x}^*$
- Option 1: Functional Approximation

$$y^* = f(\mathbf{x}^*)$$

- Option 2: Probabilistic Classifier

$$P(Y = benign|\mathbf{X} = \mathbf{x}^*), P(Y = malignant|\mathbf{X} = \mathbf{x}^*)$$

Applying Bayes Rule

Training data, $D = [\langle \mathbf{x_i}, y_i \rangle]_{i=1}^{D}$

1.
2.
3.
4.
5.

- $\mathbf{x}^* =$ `circular,small, light,irregular,thin`
- What is $P(Y = benign|\mathbf{x}^*)$?
- What is $P(Y = malignant|\mathbf{x}^*)$?

Output Label – A Discrete Random Variable

- Y takes two values
- What is $p(Y)$?
  - $\sim Ber(\theta)$
  - How do you estimate θ?
  - Treat the labels in training data as binary samples
  - Done that last week!
  - Posterior for θ

$$p(\theta) = \frac{\alpha_0 + N_1}{\alpha_0 + \beta_0 + N}$$

  - *Class 1 - Malignant; Class 2 - Benign*
  - Can we just use $p(y|\theta)$ for predicting future labels?
    - Just a prior for Y

Computing Posterior for Y

- What is probability of $\mathbf{x}^*$ to be malignant
  - $P(\mathbf{X} = \mathbf{x}^* | Y = \mathit{malignant})$?

Computing Posterior for Y

- What is probability of $\mathbf{x}^*$ to be malignant
  - $P(\mathbf{X} = \mathbf{x}^* | Y = malignant)$?
  - $P(Y = malignant)$?

Computing Posterior for Y

- What is probability of $\mathbf{x}^*$ to be malignant
  - $P(\mathbf{X} = \mathbf{x}^* | Y = \mathit{malignant})$?
  - $P(Y = \mathit{malignant})$?
  - $P(Y = \mathit{malignant} | \mathbf{X} = \mathbf{x}^*)$?

Computing Posterior for Y

- What is probability of $\mathbf{x}^*$ to be malignant
  - $P(\mathbf{X} = \mathbf{x}^* | Y = malignant)$?
  - $P(Y = malignant)$?
  - $P(Y = malignant | \mathbf{X} = \mathbf{x}^*)$?
  - $P(Y = malignant | \mathbf{X} = \mathbf{x}^*) = \frac{P(\mathbf{X}=\mathbf{x}^* | Y=malignant)P(Y=malignant)}{P(\mathbf{X}=\mathbf{x}^* | Y=malignant)P(Y=malignant)+P(\mathbf{X}=\mathbf{x}^* | Y=benign)P(Y=benign)}$

What is $P(\mathbf{X} = \mathbf{x}^* | Y = malignant)$?

- Class conditional probability of random variable $\mathbf{X}$
- **Step 1**: Assume a probability distribution for $\mathbf{X}$ ($p(\mathbf{X})$)
- **Step 2**: Learn parameters from training data

What is $P(\mathbf{X} = \mathbf{x}^* | Y = \mathit{malignant})$?

- Class conditional probability of random variable $\mathbf{X}$
- **Step 1**: Assume a probability distribution for $\mathbf{X}$ ($p(\mathbf{X})$)
- **Step 2**: Learn parameters from training data
- But $\mathbf{X}$ is multivariate discrete random variable!
- How many parameters are needed?

What is $P(\mathbf{X} = \mathbf{x}^* | Y = malignant)$?

- Class conditional probability of random variable $\mathbf{X}$
- **Step 1**: Assume a probability distribution for $\mathbf{X}$ ($p(\mathbf{X})$)
- **Step 2**: Learn parameters from training data
- But $\mathbf{X}$ is multivariate discrete random variable!
- How many parameters are needed?
- $2(2^D - 1)$

What is $P(\mathbf{X} = \mathbf{x}^* | Y = \mathit{malignant})$?

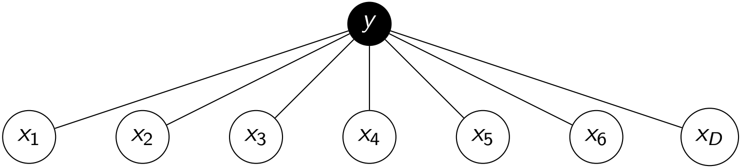
- Class conditional probability of random variable $\mathbf{X}$
- **Step 1**: Assume a probability distribution for $\mathbf{X}$ ($p(\mathbf{X})$)
- **Step 2**: Learn parameters from training data
- But $\mathbf{X}$ is multivariate discrete random variable!
- How many parameters are needed?
- $2(2^D - 1)$
- How much training data is needed?

Naive Bayes Assumption

- **All features are independent**
- Each variable can be assumed to be a Bernoulli random variable

$$P(\mathbf{X} = \mathbf{x}^*|Y = malignant) = \prod_{j=1}^{D} P(X_j = x_j^*|Y = malignant)$$

$$P(\mathbf{X} = \mathbf{x}^*|Y = benign) = \prod_{j=1}^{D} P(X_j = x_j^*|Y = benign)$$



- Only need $2D$ parameters

Example - Only binary features

- **Training a Naive Bayes Classifier**
- Find parameters that maximize likelihood of training data
  - What is a training example?
    - ~~$\mathbf{x}_i$?~~
    - $\langle \mathbf{x_i}, y_i \rangle$
  - What are the parameters?
    - θ for Y (*class prior*)
    - $\boldsymbol{\theta}_{benign}$ and $\boldsymbol{\theta}_{malignant}$ (or $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$)
  - Joint probability distribution of (X, Y)

$$\begin{aligned} p(\mathbf{x}_i, y_i) &= p(y_i|\theta)p(\mathbf{x}_i|y_i) \\ &= p(y_i|\theta)\prod_j p(x_{ij}|\theta_{jy_i}) \end{aligned}$$

Likelihood?

- Likelihood for D

$$l(D|\boldsymbol{\Theta}) = \prod_i \left(p(y_i|\theta) \prod_j p(x_{ij}|\theta_{jy_i}) \right)$$

- Log-likelihood for D

$$\begin{aligned} ll(D|\boldsymbol{\Theta}) &= N_1 \log \theta + N_2 \log(1-\theta) \\ &+ \sum_j N_{1j} \log \theta_{1j} + (N_1 - N_{1j}) \log (1 - \theta_{1j}) \\ &+ \sum_j N_{2j} \log \theta_{2j} + (N_2 - N_{2j}) \log (1 - \theta_{2j}) \end{aligned}$$

- N_1 - # malignant training examples, $N_2 =$ # benign training examples
- N_{1j} - # malignant training examples with $x_j = 1$, $N_{2j} =$ # benign training examples with $x_j = 2$

MLE?

- Maximize with respect to θ, assuming Y to be *Bernoulli*

$$\hat{\theta}_c = \frac{N_c}{N}$$

- Assuming each feature is binary ($x_j|(y = c) \sim$ *Bernoulli*(θ_{cj}), $c = \{1, 2\}$)

$$\hat{\theta}_{cj} = \frac{N_{cj}}{N_c}$$

Algorithm 1 Naive Bayes Training for Binary Features

$N_c = 0,\ N_{cj} = 0, \forall j$
for $i = 1 : N$ **do**
    $c \leftarrow y_i$
    $N_c \leftarrow N_c + 1$
    for $j = 1 : D$ **do**
        if $x_{ij} = 1$ **then**
            $N_{cj} \leftarrow N_{cj} + 1$
        end if
    end for
end for
$\hat{\theta}_c = \frac{N_c}{N},\ \hat{\theta}_{cj} = \frac{N_{cj}}{N_c}$
return b

Adding Prior

- Add prior to θ and each θ_{cj}.
  - Beta prior for θ ($\sim Beta(a_0, b_0)$)
  - Beta prior for θ_{cj} ($\sim Beta(a, b)$)

Posterior Estimates

$$p(\theta|D) = Beta(N_1 + a_0, N - N_1 + b_0)$$
$$p(\theta_{cj}|D) = Beta(N_{cj} + a, N_c - N_{cj} + b)$$

Using Naive Bayes Model for Prediction

$$p(y = c|\mathbf{x}^*, D) \propto p(y = c|D) \prod_j p(x_j^*|y = c, D)$$

- MLE approach, MAP approach?
- Bayesian approach:

$$\begin{aligned} p(y = 1|\mathbf{x}, D) \quad \propto \quad & \left[\int Ber(y = 1|\theta)p(\theta|D)d\theta\right] \\ & \prod_j \left[\int Ber(x_j|\theta_{cj})p(\theta_{cj}|D)d\theta_{cj}\right] \end{aligned}$$

$$\bar{\theta} = \frac{N_1 + a_0}{N + a_0 + b_0}$$

$$\bar{\theta}_{cj} = \frac{N_{cj} + a}{N_c + a + b}$$

Example

#	Shape	Size	Color	Type
1	cir	large	light	malignant
2	cir	large	light	benign
3	cir	large	light	malignant
4	ovl	large	light	benign
5	ovl	large	dark	malignant
6	ovl	small	dark	benign
7	ovl	small	dark	malignant
8	ovl	small	light	benign
9	cir	small	dark	benign
10	cir	large	dark	malignant

- Test example: $\mathbf{x}^* = \{cir, small, light\}$

References