Introduction to Machine Learning

Bayesian Classification

Varun Chandola

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

Learning Probabilistic Classifiers
- Treating Output Label $Y$ as a Random Variable
- Computing Posterior for $Y$
- Computing Class Conditional Probabilities

Naive Bayes Classification
- Naive Bayes Assumption
- Maximizing Likelihood
- Maximum Likelihood Estimates
- Adding Prior
- Using Naive Bayes Model for Prediction
- Naive Bayes Example

Gaussian Discriminant Analysis
- Moving to Continuous Data
- Quadratic and Linear Discriminant Analysis
- Training a QDA or LDA Classifier
Learning Probabilistic Classifiers

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

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 $x^*$

▶ **Option 1: Functional Approximation**

$$y^* = f(x^*)$$

▶ **Option 2: Probabilistic Classifier**

$$P(Y = benign | X = x^*), P(Y = malignant | X = x^*)$$
Applying Bayes Rule

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

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, irregular, thin\} - benign

- What is \( P(Y = \text{benign} | x^*) \)?
- What is \( P(Y = \text{malignant} | x^*) \)?
Output Label – A Discrete Random Variable

- $Y$ takes two values
- What is $p(Y)$?
  - $\sim \text{Ber} (\theta)$
  - How do you estimate $\theta$?
  - Treat the labels in training data as binary samples
  - Done that last week!
  - Posterior for $\theta$
    \[
    p(\theta) = \frac{\alpha_0 + N_1}{\alpha_0 + \beta_0 + N}
    \]
- Class 1 - Malignant; Class 2 - Benign
- Can we just use $p(y|\theta)$ for predicting future labels?
  - Just a prior for $Y$
What is probability of $x^*$ to be malignant

$P(X = x^* | Y = malignant)$?
Computing Posterior for $Y$

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

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

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

- Class conditional probability of random variable $X$
- **Step 1**: Assume a probability distribution for $X$ \( p(X) \)
- **Step 2**: Learn parameters from training data
What is $P(X = x^* | Y = \text{malignant})$?

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

- 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 $P(X = x^* | Y = \text{malignant})$?

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

\[
P(X = x^* | Y = \text{malignant}) = \prod_{j=1}^{D} p(x_j^* | Y = \text{malignant})
\]

\[
P(X = x^* | Y = \text{benign}) = \prod_{j=1}^{D} p(x_j^* | Y = \text{benign})
\]

- Only need 2D parameters
Training a Naive Bayes Classifier

Find parameters that maximize likelihood of training data

What is a training example?

- $x_i$?
- $\langle x_i, y_i \rangle$

What are the parameters?

- $\theta$ for $Y$ (class prior)
- $\theta_{\text{benign}}$ and $\theta_{\text{malignant}}$ (or $\theta_1$ and $\theta_2$)

Joint probability distribution of $(X, Y)$

$$p(x_i, y_i) = p(y_i|\theta)p(x_i|y_i)$$

$$= p(y_i|\theta) \prod_j p(x_{ij}|\theta_{jy_i})$$
Likelihood?

- Likelihood for $D$

\[
I(D|\Theta) = \prod_i \left( p(y_i|\theta) \prod_j p(x_{ij}|\theta_{jy_i}) \right)
\]

- Log-likelihood for $D$

\[
II(D|\Theta) = N_1 \log \theta + N_2 \log(1 - \theta) \\
+ \sum_j N_{1j} \log \theta_{1j} + (N_1 - N_{1j}) \log (1 - \theta_{1j}) \\
+ \sum_j N_{2j} \log \theta_{2j} + (N_2 - N_{2j}) \log (1 - \theta_{2j})
\]

- $N_1$ - # malignant training examples, $N_2$ = # benign training examples
- $N_{1j}$ - # malignant training examples with $x_j = 1$, $N_{2j}$ = # benign training examples with $x_j = 2$
Maximize with respect to $\theta$, assuming $Y$ to be Bernoulli

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

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

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

**Algorithm 1 Naive Bayes Training for Binary Features**

1. $N_c = 0$, $N_{cj} = 0$, $\forall j$
2. for $i = 1 : N$ do
3. $c \leftarrow y_i$
4. $N_c \leftarrow N_c + 1$
5. for $j = 1 : D$ do
6. if $x_{ij} = 1$ then
7. $N_{cj} \leftarrow N_{cj} + 1$
8. end if
9. end for
10. end for
11. $\hat{\theta}_c = \frac{N_c}{N}$, $\hat{\theta}_{cj} = \frac{N_{cj}}{N_c}$
12. return $b$
Adding Prior

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

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

\[ p(y = c|x^*, D) \propto p(y = c|D) \prod_j p(x_j^*|y = c, D) \]

- MLE approach, MAP approach?
- Bayesian approach:

\[ p(y = 1|x, D) \propto \left[ \int Ber(y = 1|\theta)p(\theta|D)d\theta \right] \prod_j \left[ \int Ber(x_j|\theta_{cj})p(\theta_{cj}|D)d\theta_{cj} \right] \]

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

\[ \bar{\theta}_{cj} = \frac{N_{cj} + a}{N_c + a + b} \]
Example

<table>
<thead>
<tr>
<th>#</th>
<th>Shape</th>
<th>Size</th>
<th>Color</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cir</td>
<td>large</td>
<td>light</td>
<td>malignant</td>
</tr>
<tr>
<td>2</td>
<td>cir</td>
<td>large</td>
<td>light</td>
<td>benign</td>
</tr>
<tr>
<td>3</td>
<td>cir</td>
<td>large</td>
<td>light</td>
<td>malignant</td>
</tr>
<tr>
<td>4</td>
<td>ovl</td>
<td>large</td>
<td>light</td>
<td>malignant</td>
</tr>
<tr>
<td>5</td>
<td>ovl</td>
<td>large</td>
<td>dark</td>
<td>benign</td>
</tr>
<tr>
<td>6</td>
<td>ovl</td>
<td>small</td>
<td>dark</td>
<td>benign</td>
</tr>
<tr>
<td>7</td>
<td>ovl</td>
<td>small</td>
<td>dark</td>
<td>malignant</td>
</tr>
<tr>
<td>8</td>
<td>ovl</td>
<td>small</td>
<td>light</td>
<td>benign</td>
</tr>
<tr>
<td>9</td>
<td>cir</td>
<td>small</td>
<td>dark</td>
<td>benign</td>
</tr>
<tr>
<td>10</td>
<td>cir</td>
<td>large</td>
<td>dark</td>
<td>malignant</td>
</tr>
</tbody>
</table>

Test example: $x^* = \{\text{cir}, \text{small}, \text{light}\}$
What if Attributes are Continuous?

- Naive Bayes is still applicable!
- Each variable is a univariate Gaussian (normal) distribution

\[
p(y|x) = p(y) \prod_{j} p(x_j|y) = p(y) \prod_{j} \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}}
\]

\[
= p(y) \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} e^{-\frac{(x-\mu)^\top \Sigma^{-1} (x-\mu)}{2}}
\]

- Where \(\Sigma\) is a diagonal matrix with \(\sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2\) as the diagonal entries
- \(\mu\) is a vector of means
- Treating \(x\) as a multivariate Gaussian with zero covariance
What if $\Sigma$ is not diagonal?

- Gaussian Discriminant Analysis
  - Class conditional density
    
    $$p(x|y = 1) = \mathcal{N}(\mu_1, \Sigma_1)$$
    
    $$p(x|y = 2) = \mathcal{N}(\mu_2, \Sigma_2)$$

- Posterior density for $y$
  
  $$p(y = 1|x) = \frac{p(y = 1)\mathcal{N}(\mu_1, \Sigma_1)}{p(y = 1)\mathcal{N}(\mu_1, \Sigma_1) + p(y = 2)\mathcal{N}(\mu_2, \Sigma_2)}$$
Using non-diagonal covariance matrices for each class - **Quadratic Discriminant Analysis (QDA)**
- Quadratic decision boundary

- If $\Sigma_1 = \Sigma_2 = \Sigma$
- **Linear Discriminant Analysis (LDA)**
  - Parameter *sharing* or *tying*
  - Results in linear surface
  - No quadratic term
Alternative Interpretation of LDA

- Equivalent to computing the **Mahalanobis distance** of \( x \) to the two means.
How to Train

MLE Training

- Estimate Bernoulli parameters for $Y$ using MLE
- For each class, estimate MLE parameters for the multivariate normal distribution, i.e., $\mu_1, \Sigma_1$ and $\mu_2, \Sigma_2$
- For LDA, compute the MLE for $\Sigma$ using all training data (ignoring the class label)