

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 Engineering and Applied Sciences



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, $D = [\langle \mathbf{x}_i, y_i \rangle]_{i=1}^D$

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

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

Applying Bayes Rule

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

1. 
2. 
3. 
4. 
5. 

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

Output Label – A Discrete Random Variable

- ▶ Y takes two values
- ▶ What is $p(Y)$?
 - ▶ $\sim \text{Ber}(\theta)$
 - ▶ 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
 - ▶ $P(\mathbf{X} = \mathbf{x}^* | Y = \textit{malignant})?$

Computing Posterior for Y

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

Computing Posterior for Y

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

Computing Posterior for Y

- ▶ What is probability of \mathbf{x}^* to be malignant

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

- ▶ $P(Y = \text{malignant})?$

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

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

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

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

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

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

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

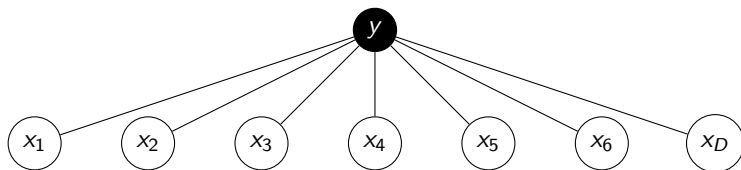
- ▶ Class conditional probability of random variable \mathbf{X}
- ▶ **Step 1:** Assume a probability distribution for \mathbf{X} ($p(\mathbf{X})$)
- ▶ **Step 2:** Learn parameters from training data
- ▶ But \mathbf{X} is multivariate discrete random variable!
- ▶ How many parameters are needed?
- ▶ $2(2^D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

Naive Bayes Assumption

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

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

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



- ▶ 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*)
 - ▶ θ_{benign} and $\theta_{malignant}$ (or θ_1 and θ_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

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

- Log-likelihood for D

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

- 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 $x_j = 2$

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

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

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

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

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  $b$ 

```

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

Posterior Estimates

$$p(\theta|D) = \text{Beta}(N_1 + a_0, N - N_1 + b_0)$$

$$p(\theta_{cj}|D) = \text{Beta}(N_{cj} + a, N_c - N_{cj} + b)$$

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:

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

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

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

Example

#	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: $\mathbf{x}^* = \{cir, small, light\}$

What if Attributes are Continuous?

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

$$\begin{aligned} p(y|\mathbf{x}) &= p(y) \prod_j p(x_j|y) = p(y) \prod_j \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}} \\ &= p(y) \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}} \end{aligned}$$

- ▶ Where Σ is a diagonal matrix with $\sigma_1^2, \sigma_1^2, \dots, \sigma_D^2$ as the diagonal entries
- ▶ $\boldsymbol{\mu}$ is a vector of means
- ▶ Treating \mathbf{x} as a multivariate Gaussian with zero covariance

What if Σ is not diagonal?

- ▶ Gaussian Discriminant Analysis

- ▶ Class conditional density

$$p(\mathbf{x}|y = 1) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

$$p(\mathbf{x}|y = 2) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

- ▶ Posterior density for y

$$p(y = 1|\mathbf{x}) = \frac{p(y = 1)\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}{p(y = 1)\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + p(y = 2)\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)}$$

Quadratic and Linear Discriminant Analysis

- ▶ 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 \mathbf{x} to the two means.

MLE Training

- ▶ Estimate Bernoulli parameters for Y using MLE
- ▶ For each class, estimate MLE parameters for the multivariate normal distribution, i.e., μ_1, Σ_1 and μ_2, Σ_2
- ▶ For LDA, compute the MLE for Σ using all training data (ignoring the class label)

References