Introduction to Machine Learning

Bayesian Classification

Mingchen Gao

September 12, 2021

Outline

Contents

1 Learning Probabilistic Classifiers

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

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

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

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

Turns out that if we have not observed the training data, then the best probabilistic estimates we can provide is $P(Y = benign) = P(Y = malignant) = 0.5$. But if we know how many times Y takes each value in a randomly sampled data set, we can make a better estimate.

1.1 Treating Output Label Y as a 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
  - 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

1.2 Computing Posterior for Y

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

1.3 Computing Class Conditional Probabilities

- 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?

Note that the $\mathbf{X}$ can take 2^D values. That means that the probability distribution should consist of probability of observing each possibility. Given that all probabilities sum to 1, we need $2^D - 1$ probabilities. We need these probabilities for each value of Y, hence $2(2^D - 1)$ probabilities.

Obviously, to reliably estimate the probabilities, one need to observe each possible realization of $\mathbf{X}$ at least a few times. Which means that we need large amounts of training data!

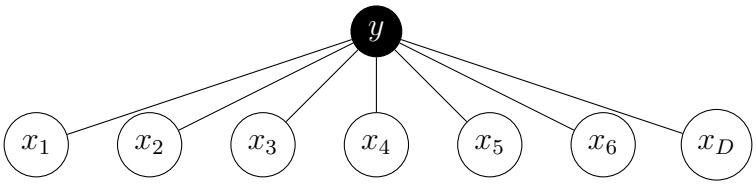
2 Naive Bayes Classification

2.1 Naive Bayes Assumption

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

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

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



- Only need $2D$ parameters
- **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)

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

2.2 Maximizing 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

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

- 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 $x_j = 2$

Derivation of the log-likelihood can be done by using the following results. The summation $\sum_i \log p(y_i|\theta)$ can be expanded and reordered by each class. For each class, the contribution to the sum will be $N_c p(y_i|\theta_c)$ where N_c is the number of training examples with c as the class label and θ_c is the class prior for class c. The double summation $\sum_i \sum_j \log p(x_{ij}|\theta_{jy_i})$ is same as $\sum_j \sum_i \log p(x_{ij}|\theta_{jy_i})$. The inner sum can be expanded and order by each class. For each class, the contribution to the sum will be $\sum_{i:y_i=c} \log p(x_{ij}|\theta_{jc})$.

2.3 Maximum Likelihood Estimates

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

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

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

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

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

2.4 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

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

2.5 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:

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

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

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

Obviously, the MLE and MAP approach use the MLE and MAP estimates of the parameters to compute the above probability.

2.6 Naive Bayes 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: $\mathbf{x}^* = \{cir, small, light\}$

We can predict a label in three ways. First is to use the MLE for all the parameters. Second is to use MAP and third is to use the Bayesian averaging approach. In each, we need to plug in the parameter estimates in:

$$\begin{aligned} P(Y = malignant|X = \mathbf{x}^*) &= \hat{\theta} \times \hat{\theta}_{malignant,cir} \times \hat{\theta}_{malignant,small} \times \hat{\theta}_{malignant,light} \\ P(Y = benign|X = \mathbf{x}^*) &= \hat{\theta} \times \hat{\theta}_{benign,cir} \times \hat{\theta}_{benign,small} \times \hat{\theta}_{benign,light} \end{aligned}$$

3 Gaussian Discriminant Analysis

3.1 Moving to Continuous Data

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

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

- Where $\mathbf{\Sigma}$ is a diagonal matrix with $\sigma_1^2, \sigma_1^2, \ldots, \sigma_D^2$ as the diagonal entries
- $\boldsymbol{\mu}$ is a vector of means
- Treating $\mathbf{x}$ as a multivariate Gaussian with zero covariance
- Gaussian Discriminant Analysis
  - Class conditional density

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

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

  - Posterior density for y

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

3.2 Quadratic and Linear Discriminant Analysis

- Using non-diagonal covariance matrices for each class - **Quadratic Discriminant Analysis (QDA)**
  - Quadratic decision boundary
- If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$
- **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.
- **Euclidean distance** is a special case of Mahalanobis distance when Σ is an identity matrix.

One can geometrically interpret the Linear Discriminant Analysis by noting that the exponential in the *pdf* of a multivariate gaussian:

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

is the **Mahalanobis Distance** between an example $\mathbf{x}$ and the mean $\boldsymbol{\mu}$ in the D dimensional space.

3.3 Training a QDA or LDA Classifier

MLE Training

- Estimate Bernoulli parameters for Y using MLE
- For each class, estimate MLE parameters for the multivariate normal distribution, i.e., $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$
- For LDA, compute the MLE for $\boldsymbol{\Sigma}$ using all training data (ignoring the class label)

References

Murphy Book Chapters 3.5 4.2