Generative Model (Naïve Bayes, LDA)

IST557 Data Mining: Techniques and Applications
Jessie Li, Penn State University
Review: Bayesian Statistics

posterior

\[ p(\theta | D) \propto p(D | \theta) p(\theta) \]

likelihood

prior

Final guess

Observe 3 heads, 2 tails, guess 3/5

Head, tail probability is close to 0.5 and 0.5

Beta(\( \beta_H + \alpha_H, \beta_T + \alpha_T \))

\[ \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \]

Beta(\( \beta_H, \beta_T \))

Maximum a Posterior (MAP)

Maximum Likelihood (MLE)

When uniform prior, MLE=MAP
Review: Bayesian Statistics

\[ p(\theta | D) \propto p(D | \theta) p(\theta) \]

- **Posterior**
- **Likelihood**
- **Prior**

Maximum a Posterior (MAP)

**Discriminative** model:
- Logistic regression
- SVM
- Decision tree
- KNN

Maximum Likelihood (MLE)

**Generative** model:
- Naïve bayes classifier
- Linear discriminant analysis
- Mixture model
- Hidden Markov Model
Review: Bayesian Rule

\[ p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \]

Different notations

\[ p(y|x) = \frac{p(x|y)p(y)}{p(x)} \]

in most materials

\[ p(G|X) = \frac{p(X|G)p(G)}{p(X)} \]

In today’s lecture

D: data
\( \theta \): model
Naïve Bayes Classifier
Naïve Bayes Classifier

• A “naïve” assumption
  – attributes are conditionally independent (i.e., no dependence relation between attributes)

\[
p(X|G) = p(x_1|G)p(x_2|G) \cdots p(x_k|G)
\]
Naïve Bayes Classifier: Example

Class:
buys_computer = ‘yes’
buys_computer = ‘no’

Data to be classified:
X = (age <=30,
Income = medium,
Student = yes
Credit_rating = Fair)
Naïve Bayes Classifier: Example

**Class:** ‘yes’ or ‘no’

**Data to be classified**

\[ X = (\text{age} \leq 30, \text{Income} = \text{medium}, \text{Student} = \text{yes}, \text{Credit\_rating} = \text{Fair}) \]

\[
\begin{align*}
\Pr(\text{yes}|X) & \propto \Pr(X|\text{yes})\Pr(\text{yes}) \\
\Pr(X|\text{yes}) & = \Pr(\text{age} \leq 30|\text{yes}) \times \Pr(\text{income} = \text{medium}|\text{yes}) \\
& \quad \times \Pr(\text{student} = \text{yes}|\text{yes}) \times \Pr(\text{credit} = \text{fair}|\text{yes})
\end{align*}
\]

\[
\begin{align*}
\Pr(\text{no}|X) & \propto \Pr(X|\text{no})\Pr(\text{no})
\end{align*}
\]

**Prediction:** \( \Pr(\text{yes}|X) > \Pr(\text{no}|X) \)?
Naïve Bayes Classifier: Example

Data: \( X = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair}) \)

**Prior**
P(\( \text{yes} \)) = 9/14 = 0.643

**Likelihood**
P(\( \text{age} = \text{“<=30”} \mid \text{yes} \)) = 2/9 = 0.222

P(\( \text{income} = \text{“medium”} \mid \text{yes} \)) = 4/9 = 0.444

P(\( \text{student} = \text{“yes”} \mid \text{yes} \)) = 6/9 = 0.667

P(\( \text{credit\_rating} = \text{“fair”} \mid \text{yes} \)) = 6/9 = 0.667

P(\( X \mid \text{yes} \)) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044

**Posterior**
P(\( X \mid \text{yes} \)) \times P(\( \text{yes} \)) = 0.044 \times 0.643 = 0.028
Naïve Bayes Classifier: Example

Data: X = (age <= 30, income = medium, student = yes, credit_rating = fair)

Prior
P(no) = 5/14 = 0.357

Likelihood
P(age = “<=30” | no) = 3/5 = 0.6
P(income = “medium” | no) = 2/5 = 0.4
P(student = “yes” | no) = 1/5 = 0.2
P(credit_rating = “fair” | no) = 2/5 = 0.4

P(X | no) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

Posterior
P(X | no) * P(no) = 0.019 x 0.357 = 0.007
Naïve Bayes Classifier: Example

Data: $X = (\text{age} \leq 30, \ \text{income} = \text{medium}, \ \text{student} = \text{yes}, \ \text{credit\_rating} = \text{fair})$

**Posterior**

$P(\text{yes} | X) = 0.028$

$P(\text{no} | X) = 0.007$

$P(\text{yes} | X) > P(\text{no} | X) \Rightarrow \text{Yes}$
Avoiding the Zero-Probability Problem

• Naïve Bayesian prediction requires each conditional prob. be non-zero. Otherwise, the predicted prob. will be zero

\[ p(X|G) = p(x_1|G)p(x_2|G) \cdots p(x_k|G) \]

• Ex. Suppose a dataset with 1000 tuples, income=low (0), income=medium (990), and income = high (10)

• Use Laplacian correction (or Laplacian estimator)
  – Adding 1 to each case
    Prob(income = low) = 1/1003
    Prob(income = medium) = 991/1003
    Prob(income = high) = 11/1003
  – The “corrected” prob. estimates are close to their “uncorrected” counterparts
Naïve Bayes Classifier: Continuous Example

Class: Male or Female
Data:
X= (height=6, weight=130, foot=8)

P(male | X) = P(X | male) P(male)
= P(height=6 | male)P(weight=130 | male)P(foot=8 | male) P(male)

P(male) = 0.5
P(height=6 | male)=???
Naïve Bayes Classifier: Continuous Example

Class: Male or Female
Data:
\[ X = (\text{height}=6, \text{weight}=130, \text{foot}=8) \]

\[ P(\text{height}=6|\text{male})=??? \]

Use Gaussian distribution for continuous data estimation

\[ p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma^2_c}} e^{-\frac{(v-\mu_c)^2}{2\sigma^2_c}} \]

<table>
<thead>
<tr>
<th>sex</th>
<th>mean (height)</th>
<th>variance (height)</th>
<th>mean (weight)</th>
<th>variance (weight)</th>
<th>mean (foot size)</th>
<th>variance (foot size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>5.855</td>
<td>3.50E-02</td>
<td>176.25</td>
<td>1.23E+02</td>
<td>11.25</td>
<td>9.17E-01</td>
</tr>
<tr>
<td>female</td>
<td>5.4175</td>
<td>9.72E-02</td>
<td>132.5</td>
<td>5.58E+02</td>
<td>7.5</td>
<td>1.67E+00</td>
</tr>
</tbody>
</table>

\[ P(\text{height}=6|\text{male}) = 1.5789 \]

Source: wikipedia
Naïve Bayes Classifier: 
Continuous Example

Class: Male or Female
Data: \(X = (\text{height}=6, \text{weight}=130, \text{foot}=8)\)

\[
P(\text{male}) = 0.5
\]

\[
p(\text{height}|\text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6 - \mu)^2}{2\sigma^2}\right) \approx 1.5789
\]

\[
p(\text{weight}|\text{male}) = 5.9881 \cdot 10^{-6}
\]

\[
p(\text{foot size}|\text{male}) = 1.3112 \cdot 10^{-3}
\]

posterior numerator (male) = their product = \(6.1984 \cdot 10^{-9}\)

\[
P(\text{female}) = 0.5
\]

\[
p(\text{height}|\text{female}) = 2.2346 \cdot 10^{-1}
\]

\[
p(\text{weight}|\text{female}) = 1.6789 \cdot 10^{-2}
\]

\[
p(\text{foot size}|\text{female}) = 2.8669 \cdot 10^{-1}
\]

posterior numerator (female) = their product = \(5.3778 \cdot 10^{-4}\)

Source: wikipedia
Linear/Quadratic Discriminant Analysis (LDA/QDA)
Linear Discriminant Analysis (LDA)

\[ Pr(G = k|X = x) \propto Pr(X = x|G = k)Pr(G = k) \]

\[ f_k(x) \quad \pi_k \]

Denote as

- **Multivariable Gaussian**
  
  \[ f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)} \]

- Formula above is Quadratic Discriminant Analysis (QDA)
- Linear Discriminant Analysis (LDA): \[ \Sigma_k = \Sigma, \forall k. \]

- **LDA**
  - Mean defines place, covariance defines the shape
  - Same shapes across all classes (shifted versions)
LDA: Class Density

3 classes. same shapes, different mean
LDA: Class Density
Solve Optimization

- **Optimal classification**

\[
f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}
\]

\[
\hat{G}(x) = \arg \max_k \Pr(G = k \mid X = x)
\]

\[
= \arg \max_k f_k(x) \pi_k = \arg \max_k \log(f_k(x) \pi_k)
\]

\[
= \arg \max_k \left[ -\log((2\pi)^{p/2} |\Sigma|^{1/2}) -\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k) + \log(\pi_k) \right]
\]

\[
= \arg \max_k \left[ -\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k) + \log(\pi_k) \right]
\]

**Note**

\[
-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k - \frac{1}{2} x^T \Sigma^{-1} x
\]
Solve Optimization

To sum up

\[ \hat{G}(x) = \arg \max_k \left[ x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \right] \]

- Define the linear discriminant function

\[ \delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k). \]

Then

\[ \hat{G}(x) = \arg \max_k \delta_k(x). \]

- The decision boundary between class \( k \) and \( l \) is:

\[ \left\{ x : \delta_k(x) = \delta_l(x) \right\}. \]

Or equivalently the following holds

\[ \log \frac{\pi_k}{\pi_l} - \frac{1}{2} (\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l) = 0. \]
Example: Binary Classification

Binary classification ($k = 1$, $l = 2$):

- Define $a_0 = \log \frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1 + \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)$.
- Define $(a_1, a_2, ..., a_p)^T = \Sigma^{-1}(\mu_1 - \mu_2)$.
- Classify to class 1 if $a_0 + \sum_{j=1}^{p} a_j x_j > 0$; to class 2 otherwise.
- An example:
  - $\pi_1 = \pi_2 = 0.5$.
  - $\mu_1 = (0, 0)^T$, $\mu_2 = (2, -2)^T$.
  - $\Sigma = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 0.5625 \end{pmatrix}$.
  - Decision boundary:
    $$5.56 - 2.00x_1 + 3.56x_2 = 0.0.$$
Example: Binary Classification
Murphy, Figure 4.5, discrimAnalysisDboundariesDemo
Estimate Gaussian Distribution

- In practice, we need to estimate the Gaussian distribution.
- \( \hat{\pi}_k = \frac{N_k}{N} \), where \( N_k \) is the number of class-\( k \) samples.
- \( \hat{\mu}_k = \frac{\sum_{g_i=k} x^{(i)}}{N_k} \).
- \( \hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x^{(i)} - \hat{\mu}_k)(x^{(i)} - \hat{\mu}_k)^T / (N - K) \).
- Note that \( x^{(i)} \) denotes the \( i \)th sample vector.
Diabetes Example

The diabetes data set is taken from the UCI machine learning database repository at:
http://www.ics.uci.edu/~mlearn/Machine-Learning.html. The original source of the data is the National Institute of Diabetes and Digestive and Kidney Diseases. There are 768 cases in the data set, of which 268 show signs of diabetes according to World Health Organization criteria. Each case contains 8 quantitative variables, including diastolic blood pressure, triceps skin fold thickness, a body mass index, etc.

- Two classes: with or without signs of diabetes.
- Denote the 8 original variables by \( \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_8 \).
- Remove the mean of \( \tilde{X}_j \) and normalize it to unit variance.

R script for this example: https://onlinecourses.science.psu.edu/stat857/node/135
Diabetes Example: PCA

The two principal components $X_1$ and $X_2$ are used in classification:

$$X_1 = 0.1284\tilde{X}_1 + 0.3931\tilde{X}_2 + 0.3600\tilde{X}_3 + 0.4398\tilde{X}_4$$
$$+ 0.4350\tilde{X}_5 + 0.4519\tilde{X}_6 + 0.2706\tilde{X}_7 + 0.1980\tilde{X}_8$$

$$X_2 = 0.5938\tilde{X}_1 + 0.1740\tilde{X}_2 + 0.1839\tilde{X}_3 - 0.3320\tilde{X}_4$$
$$- 0.2508\tilde{X}_5 - 0.1010\tilde{X}_6 - 0.1221\tilde{X}_7 + 0.6206\tilde{X}_8$$
PCA: Principle Component Analysis

- PCA is a mathematical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- Frequently used for dimension reduction
Diabetes Example

The scatter plot follows. Without diabetes: stars (class 1), with diabetes: circles (class 2).
Example: Diabetes Data - LDA

- Two input variables computed from the principal components of the original 8 variables.
- Prior probabilities: \( \hat{\pi}_1 = 0.651, \hat{\pi}_2 = 0.349. \)
- \( \hat{\mu}_1 = (-0.4035, -0.1935)^T, \hat{\mu}_2 = (0.7528, 0.3611)^T. \)
- \( \hat{\Sigma} = \begin{pmatrix} 1.7925 & -0.1461 \\ -0.1461 & 1.6634 \end{pmatrix} \)
- Classification rule:

\[
\hat{G}(x) = \begin{cases} 
1 & 0.7748 - 0.6771x_1 - 0.3929x_2 \geq 0 \\
2 & \text{otherwise}
\end{cases}
\]

\[
= \begin{cases} 
1 & 1.1443 - x_1 - 0.5802x_2 \geq 0 \\
2 & \text{otherwise}
\end{cases}
\]
Diabetes Example: LDA Decision Boundary

The scatter plot follows. Without diabetes: stars (class 1), with diabetes: circles (class 2). Solid line: classification boundary obtained by LDA. Dash dot line: boundary obtained by linear regression of indicator matrix.
Diabetes Example: Evaluation

- Within training data classification error rate: 28.26%.
- Sensitivity: 45.90%.
- Specificity: 85.60%

Error rate = \( \frac{\text{# of samples having wrong prediction}}{\text{# of samples}} \)

\[
\text{Sensitivity} = \frac{\text{# of samples that prediction as positive and truth is positive}}{\text{# of samples that prediction as positive}}
\]

\[
\text{Specificity} = \frac{\text{# of samples that prediction as negative and truth is negative}}{\text{# of samples that prediction as negative}}
\]
Quadratic Discriminant Analysis (QDA)

- Estimate the covariance matrix $\Sigma_k$ separately for each class $k$, $k = 1, 2, \ldots, K$.

- Quadratic discriminant function:

$$
\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k .
$$

- Classification rule:

$$
\hat{G}(x) = \arg \max_k \delta_k(x) .
$$

- Decision boundaries are quadratic equations in $x$.

- QDA fits the data better than LDA, but has more parameters to estimate.
Diabetes Example: QDA

- Prior probabilities: $\hat{\pi}_1 = 0.651$, $\hat{\pi}_2 = 0.349$.
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T$, $\hat{\mu}_2 = (0.7528, 0.3611)^T$.
- $\hat{\Sigma}_1 = \begin{pmatrix} 1.6769 & -0.0461 \\ -0.0461 & 1.5964 \end{pmatrix}$
- $\hat{\Sigma}_2 = \begin{pmatrix} 2.0087 & -0.3330 \\ -0.3330 & 1.7887 \end{pmatrix}$
Diabetes Example: QDA Decision Boundary
Diabetes Example: QDA Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Error Rate</th>
<th>Sensitivity</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>28.26%</td>
<td>45.90%</td>
<td>85.60%</td>
</tr>
<tr>
<td>QDA</td>
<td>29.04%</td>
<td>45.90%</td>
<td>84.40%</td>
</tr>
</tbody>
</table>

Sensitivity (QDA) is the same as LDA, but specification is lower
Summary

• **Generative model**
  – Assume data is generated by some model: \( P(D|\theta) \)

• **Naïve Bayes classifier**
  – Features are independent
  – Categorical features
  – Continuous features

• **Linear discriminant analysis and (LDA) and quadratic discriminant analysis (QDA)**
  – Multivariate Gaussian distribution
  – Differences between LDA and QDA
    • LDA assumes the covariance across classes are the same