# Chapter X: Classification*

1. Basic idea
2. Decision trees
3. Naïve Bayes classifier
4. Support vector machines
5. Ensemble methods

* Zaki & Meira: Ch. 24, 26, 28 & 29; Tan, Steinbach & Kumar: Ch. 4, 5.3–5.6

# X.3 Naïve Bayes classifier

1. Basic idea
2. Computing the probabilities
3. Summary

Zaki & Meira, Ch. 26; Tan, Steinbach & Kumar, Ch. 5.3

# Basic idea

- Recall the Bayes theorem

$$\Pr[Y \mid X] = \frac{\Pr[X \mid Y]\Pr[Y]}{\Pr[X]}$$

- In classification
  - RV $X$ = attribute set
  - RV $Y$ = class variable
  - $Y$ depends on $X$ in a *non-deterministic* way
- The dependency between $X$ and $Y$ is captured in $\Pr[Y \mid X]$ and $\Pr[Y]$
  - Posterior and prior probability

# Building the classifier

- **Training phase**
  - Learn the posterior probabilities $\Pr[Y \mid X]$ for every combination of $X$ and $Y$ based on training data
- **Test phase**
  - For test record $X'$, compute the class $Y'$ that *maximizes the posterior probability* $\Pr[Y' \mid X']$
    - $Y' = \arg\max_i\{\Pr[c_i \mid X']\} = \arg\max_i\{\Pr[X' \mid c_i]\Pr[c_i]/\Pr[X']\}$
      $= \arg\max_i\{\Pr[X' \mid c_i]\Pr[c_i]\}$
- So we need $\Pr[X' \mid c_i]$ and $\Pr[c_i]$
  - $\Pr[c_i]$ is the fraction of test records that belong to class $c_i$
  - $\Pr[X' \mid c_i]$?

# Computing the probabilities

- Assume that the attributes are conditionally independent given the class label
  - Naïvety of the classifier
  - $\Pr[X \mid Y = c_i] = \prod_{i=1}^{d} \Pr[X_i \mid Y = c_i]$
    - $X_i$ is the attribute $i$
- Without independency there would be too many variables to estimate
- With independency, it is enough to estimate $\Pr[X_i \mid Y]$
  - $\Pr[Y \mid X] = \Pr[Y] \prod_{i=1}^{d} \Pr[X_i \mid Y] / \Pr[X]$
  - Pr[X] is fixed, so can be omitted
- But how to estimate the *likelihood* $\Pr[X_i \mid Y]$?

# Categorical attributes

- If $X_i$ is categorical $\Pr[X_i = x_i \mid Y = c]$ is the fraction of training instances in class $c$ that take value $x_i$ on the $i$-th attribute

Pr[HomeOwner = yes | No] = 3/7
Pr[MartialStatus = S | Yes] = 2/3

| Tid | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Continuous attributes: discretization

- We can discretize continuous attributes to intervals
  - These intervals act like ordinal attributes
- Problem is where to discretize
  - Too many intervals: too few training records per interval
    $\Rightarrow$ unreliable estimates
  - Too few intervals: intervals merge attributes from different classes and don’t help distinguishing the classes

# Continuous attributes continue

- Alternatively we can assume distribution for the continuous variables
  - Normally we assume normal distribution
- We need to estimate the distribution parameters
  - For normal distribution we can use sample mean and sample variance
  - For estimation we consider the values of attribute $X_i$ that are associated with class $c_j$ in the test data
- We hope that the parameters for distributions are different for different classes of the same attribute
  - Why?

# Naïve Bayes example

| Tid | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Annual Income:

Class = No

- Sample mean = 110
- Sample variance = 2975

Class = Yes

- Sample mean = 90
- Sample variance = 25

Test data: X = (HO = No, MS = M, AI = $120K)

# Naïve Bayes example

| Tid | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Annual Income:
Class = No
- Sample mean = 110
- Sample variance = 2975

Class = Yes
- Sample mean = 90
- Sample variance = 25

Test data: X = (HO = No, MS = M, AI = $120K)

Pr[Yes] = 0.3, Pr[No] = 0.7

# Naïve Bayes example

| Tid | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Annual Income:
Class = No
    Sample mean = 110
    Sample variance = 2975
Class = Yes
    Sample mean = 90
    Sample variance = 25

Test data: X = (HO = No, MS = M, AI = \$120K)

Pr[Yes] = 0.3, Pr[No] = 0.7

Pr[X | No] = Pr[HO = No | No] × Pr[MS = M | No] × Pr[AI = \$120K | No]
= 4/7 × 4/7 × 0.0072 = 0.0024

# Naïve Bayes example

| Tid | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Annual Income:
Class = No
- Sample mean = 110
- Sample variance = 2975

Class = Yes
- Sample mean = 90
- Sample variance = 25

Test data: X = (HO = No, MS = M, AI = \$120K)

Pr[Yes] = 0.3, Pr[No] = 0.7

$$\Pr[X \mid No] = \Pr[HO = No \mid No] \times \Pr[MS = M \mid No] \times \Pr[AI = \$120K \mid No]$$
$$= 4/7 \times 4/7 \times 0.0072 = 0.0024$$

$$\Pr[X \mid Yes] = \Pr[HO = No \mid Yes] \times \Pr[MS = M \mid Yes] \times \Pr[AI = \$120K \mid Yes]$$
$$= 1 \times 0 \times \varepsilon = 0$$

# Naïve Bayes example

| Tid | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Annual Income:
Class = No
  Sample mean = 110
  Sample variance = 2975
Class = Yes
  Sample mean = 90
  Sample variance = 25

Test data: X = (HO = No, MS = M, AI = \$120K)

$\Pr[\text{Yes}] = 0.3$, $\Pr[\text{No}] = 0.7$

$$\Pr[X \mid \text{No}] = \Pr[HO = \text{No} \mid \text{No}] \times \Pr[MS = M \mid \text{No}] \times \Pr[AI = \$120K \mid \text{No}]$$
$$= 4/7 \times 4/7 \times 0.0072 = 0.0024$$

$$\Pr[X \mid \text{Yes}] = \Pr[HO = \text{No} \mid \text{Yes}] \times \Pr[MS = M \mid \text{Yes}] \times \Pr[AI = \$120K \mid \text{Yes}]$$
$$= 1 \times 0 \times \varepsilon = 0$$

$$\Pr[\text{No} \mid X] = \alpha \times 0.7 \times 0.0024 = 0.0016\alpha, \quad \alpha = 1/\Pr[X]$$

$\Rightarrow$ $\Pr[\text{No} \mid X]$ has higher posterior and X should be classified as non-defaulter

# Naïve Bayes example

| Tid | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Annual Income:
Class = No
  Sample mean = 110
  Sample variance = 2975
Class = Yes
  Sample mean = 90
  Sample variance = 25


Test data: X = (HO = No, MS = M, AI = \$120K)

Pr[Yes] = 0.3, Pr[No] = 0.7

Pr[X | No] = Pr[HO = No | No] × Pr[MS = M | No] × Pr[AI = \$120K | No]
  = 4/7 × 4/7 × 0.0072 = 0.0024

Pr[X | Yes] = Pr[HO = No | Yes] × Pr[MS = M | Yes] × Pr[AI = \$120K | Yes]
  = 1 × 0 × $\varepsilon$ = 0

Pr[No | X] = $\alpha$ × 0.7 × 0.0024 = 0.0016$\alpha$,  $\alpha$ = 1/Pr[X]

⇒ Pr[No | X] has higher posterior and X should be classified as non-defaulter

# Continuous distributions at fixed point

- If $X_i$ is continuous, $\Pr[X_i = x_i \mid Y = c_i] = 0$!
  - But we still need to estimate that number
- Self-cancelling trick:

$$\Pr[x_i - \epsilon \leqslant X_i \leqslant x_i + \epsilon \mid Y = c_j] = \int_{x_i-\epsilon}^{x_i+\epsilon} (2\pi\sigma_{ij})^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu_{ij})^2}{2\sigma_{ij}^2}\right)$$
$$\approx 2\epsilon\, f(x_i; \mu_{ij}, \sigma_{ij})$$

  - But 2ε cancels out in the normalization constant…

# Zero likelihood

- We might have no samples with $X_i = x_i$ and $Y = c_j$
  - Naturally only problem with categorical variables
  - $\Pr[X_i = x_i \mid Y = c_j] = 0 \Rightarrow$ zero posterior probability
  - It can be that *all* classes have zero posterior probability for some validation data
- Answer is smoothing (*m-estimate*):
  - $\Pr[X_i = x_i \mid Y = c_j] = \frac{n_i + mp}{n + m}$
    - $n$ = # of training instances from class $c_j$
    - $n_i$ = # training instances from $c_j$ that take value $x_i$
    - $m$ = ”equivalent sample size”
    - $p$ = user-set parameter

# More on m-estimate

$$\Pr[X_i = x_i \mid Y = c_j] = \frac{n_i + mp}{n + m}$$

- The parameters are $p$ and $m$
  - If $n = 0$, then likelihood is $p$
    - $p$ is ”prior” of observing $x_i$ in class $c_j$
  - Parameter $m$ governs the trade-off between $p$ and observed probability $n_i/n$
- Setting these parameters is again problematic…

# More on m-estimate

$$\Pr[X_i = x_i \mid Y = c_j] = \frac{n_i + mp}{n + m}$$

- The parameters are *p* and *m*
  - If $n = 0$, then likelihood is *p*
    - *p* is ”prior” of observing $x_i$ in class $c_j$
  - Parameter *m* governs the trade-off between *p* and observed probability $n_i/n$
- Setting these parameters is again problematic…
- Alternatively, we can just add one *pseudo-count* to each class
  - $\Pr[X_i = x_i \mid Y = c_j] = (n_j + 1) / (n + |\text{dom}(X_i)|)$
    - $|\text{dom}(X_i)|$ = # values attribute $X_i$ can take

# Summary of naïve Bayes

- Robust to isolated noise
  - Averaged out
- Can handle missing values
  - Example is ignored when building the model and attribute is ignored when classifying new data
- Robust to irrelevant attributes
  - $\Pr(X_i \mid Y)$ is (almost) uniform for irrelevant $X_i$
- Can have issues with correlated attributes