Today:
• Bayes Classifiers
• Naïve Bayes
• Gaussian Naïve Bayes

Readings:
Mitchell:
“Naïve Bayes and Logistic Regression”
(available on class website)

Estimating Parameters

• Maximum Likelihood Estimate (MLE): choose $\theta$ that maximizes probability of observed data $\mathcal{D}$

$$\hat{\theta} = \arg \max_{\theta} \ P(\mathcal{D} | \theta)$$

• Maximum a Posteriori (MAP) estimate: choose $\theta$ that is most probable given prior probability and the data

$$\hat{\theta} = \arg \max_{\theta} \ P(\theta | \mathcal{D})$$
$$= \arg \max_{\theta} \ = \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}$$
Conjugate priors

- $P(\theta)$ and $P(\theta | D)$ have the same form

**Eg. 1** Coin flip problem

Likelihood is $\sim$ Binomial

$$P(D | \theta) = \theta^x (1 - \theta)^{1-x}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\alpha_H - 1} (1 - \theta)^{\beta_H - 1}}{B(\alpha_H, \beta_H)}$$ $\sim$ Beta($\beta_H, \beta_T$)

Then posterior is Beta distribution

$$P(\theta | D) \sim Beta(\alpha_H + x, \beta_H + 1 - x)$$

For Binomial, conjugate prior is Beta distribution.

[A. Singh]

---

Conjugate priors

- $P(\theta)$ and $P(\theta | D)$ have the same form

**Eg. 2** Dice roll problem (6 outcomes instead of 2)

Likelihood is $\sim$ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(D | \theta) = \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1}}{B(\alpha_1, \alpha_2, \ldots, \alpha_k)} \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$$

Then posterior is Dirichlet distribution

$$P(\theta | D) \sim Dirichlet(\alpha_1 + \beta_1, \ldots, \alpha_k + \beta_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

[A. Singh]
Conjugate priors

- $P(\theta)$ and $P(\theta|D)$ have the same form

**Eg. 2** Dice roll problem (6 outcomes instead of 2)

Likelihood is $\sim$ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(D | \theta) = \theta_1^{a_1} \theta_2^{a_2} \ldots \theta_k^{a_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1-1} \theta_2^{\beta_2-1} \ldots \theta_k^{\beta_k-1}}{B(\beta_1, \beta_2, \ldots \beta_K)} \sim \text{Dirichlet}$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

---

Let's learn classifiers by learning $P(Y|X)$

Consider $Y=$Wealth, $X=$<Gender, HoursWorked>

<table>
<thead>
<tr>
<th>gender</th>
<th>hours_worked</th>
<th>wealth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0:40.5-</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
<tr>
<td>Male</td>
<td>v0:40.5-</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
</tbody>
</table>

| Gender | HrsWorked | P( rich | G,HW) | P(poor | G,HW) |
|--------|-----------|--------|--------|
| F      | <40.5     | .09    | .91    |
| F      | >40.5     | .21    | .79    |
| M      | <40.5     | .23    | .77    |
| M      | >40.5     | .38    | .62    |
How many parameters must we estimate?

Suppose \( X = \langle X_1, \ldots, X_n \rangle \)
where \( X_i \) and \( Y \) are boolean RV’s

To estimate \( P(Y | X_1, X_2, \ldots, X_n) \)
\[
2^n \text{ params}
\]

If we have 30 boolean \( X_i \)’s: \( P(Y | X_1, X_2, \ldots, X_{30}) \)
\[
2^{30} \sim 1B
\]

Bayes Rule

\[
P(Y | X) = \frac{P(X | Y) P(Y)}{P(X)}
\]

Which is shorthand for:

\[
(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}
\]

Equivalently:

\[
(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k) P(Y = y_k)}
\]
Can we reduce params using Bayes Rule?

Suppose \( X = \langle X_1, \ldots, X_n \rangle \), where \( X_i \) and \( Y \) are boolean RV's. Using Bayes Rule:

\[
P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}
\]

\[
P(X_1, \ldots, X_n | Y) \Rightarrow \prod_{i=1}^{n} P(X_i|Y)
\]

\[
\prod_{i} P(X_i=Y_i | x_{i}', x_{-i} \neq Y_i')
\]

\[
P(Y) \Rightarrow \prod_{i} P(X_i=1 | Y)
\]

Naïve Bayes

Naïve Bayes assumes:

\[
P(X_1 \ldots X_n|Y) = \prod_{i} P(X_i|Y)
\]

i.e., that \( X_i \) and \( X_j \) are conditionally independent given \( Y \), for all \( i \neq j \)
Conditional Independence

Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

\[(\forall i, j, k) P(X = x_i|Y = y_j, Z = z_k) = P(X = x_i|Z = z_k)\]

Which we often write

\[P(X|Y, Z) = P(X|Z)\]

E.g.,

\[P(\text{Thunder}|\text{Rain, Lightning}) = P(\text{Thunder}|\text{Lightning})\]

Naïve Bayes uses assumption that the \(X_i\) are conditionally independent, given Y

Given this assumption, then:

\[
P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y) = P(X_1|Y)P(X_2|Y)
\]

in general: \[P(X_1...X_n|Y) = \prod_i P(X_i|Y)\]

How many parameters to describe \(P(X_1...X_n|Y)\)? \(P(Y)\)?

- Without conditional indep assumption? \(2(2^n - 1)\)
- With conditional indep assumption? \(2n \text{ params } + 1\)

\[P(X_i|Y) \Rightarrow 2 \text{ params}\]
Naïve Bayes in a Nutshell

Bayes rule:
\[ P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)} \]

Assuming conditional independence among \(X_i\)’s:
\[ P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)} \]

So, to pick most probable \(Y\) for \(X_{new} = < X_1, \ldots, X_n >\)
\[ Y_{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_{i_{new}} | Y = y_k) \]

---

Naïve Bayes Algorithm – discrete \(X_i\)

- Train Naïve Bayes (examples)
  - for each\(^*\) value \(y_k\)
    - estimate \(\pi_k \equiv P(Y = y_k)\)
  - for each\(^*\) value \(x_{ij}\) of each attribute \(X_i\)
    - estimate \(\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)\)

- Classify \(X_{new}\)
  \[ Y_{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_{i_{new}} | Y = y_k) \]
  \[ Y_{new} \leftarrow \arg \max_{y_k} \pi_k \prod_i \theta_{ijk} \]

\(^*\) probabilities must sum to 1, so need estimate only n-1 of these...
Estimating Parameters: $Y, X_i$ discrete-valued

Maximum likelihood estimates (MLE’s):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D \{X_i = x_{ij} \land Y = y_k\}}{\#D \{Y = y_k\}}$$

Example: Live in Sq Hill? $P(S|G,D,E)$

- $S=1$ iff live in Squirrel Hill
- $G=1$ iff shop at SH Giant Eagle
- $D=1$ iff Drive to CMU
- $E=1$ iff even # of letters in last name

What probability parameters must we estimate?
Example: Live in Sq Hill? $P(S|G,D,E)$

- $S=1$ iff live in Squirrel Hill
- $G=1$ iff shop at SH Giant Eagle
- $D=1$ iff Drive or Carpool to CMU
- $E=1$ iff Even # letters last name

| $P(S=1)$ | 26/110 |
| $P(D=1 | S=1)$ | 2/26 |
| $P(D=1 | S=0)$ | 2/84 |
| $P(G=1 | S=1)$ | 24/26 |
| $P(G=1 | S=0)$ | 14/84 |
| $P(E=1 | S=1)$ | 12/26 |
| $P(E=1 | S=0)$ | 30/84 |

$P(S=1 | G=1, D=0, E=1) = \frac{P(G=1 | S=1) \cdot P(D=0 | S=1) \cdot P(E=1 | S=1) \cdot P(S=1)}{P(G=1 | S=0) \cdot P(D=0 | S=0) \cdot P(E=1 | S=0) \cdot P(S=0)}$

Naïve Bayes: Subtlety #1

If unlucky, our MLE estimate for $P(X_i | Y)$ might be zero. (e.g., $X_i =$ Birthday_Is_January_30_1990)

- Why worry about just one parameter out of many?

- What can be done to avoid this?
Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose $\theta$ that maximizes probability of observed data $\mathcal{D}$

$$\hat{\theta} = \arg \max_{\theta} P(\mathcal{D} | \theta)$$

- Maximum a Posteriori (MAP) estimate: choose $\theta$ that is most probable given prior probability and the data

$$\hat{\theta} = \arg \max_{\theta} P(\theta | \mathcal{D})$$

$$= \arg \max_{\theta} \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}$$

Estimating Parameters: $Y, X_i$ discrete-valued

Maximum likelihood estimates:

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \land Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates (Beta, Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m(\beta_m - 1)}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \land Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m(\beta_m - 1)}$$

Only difference: “imaginary” examples
Naïve Bayes: Subtlety #2

Often the $X_i$ are not really conditionally independent

• We use Naïve Bayes in many cases anyway, and it often works pretty well
  – often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

• What is effect on estimated $P(Y|X)$?
  – Special case: what if we add two copies: $X_i = X_k$