Recitation 3

Naive Bayes and Logistic Regression

and a surprise...

Ekaterina Spriggs, 10701/15781 Fall 2009
Recitation 3

Naive Bayes and Logistic Regression

and decision trees!

Ekaterina Spriggs, 10701/15781 Fall 2009
Classifiers

\[ P(Y|X_1, \ldots, X_n) \]

**Directly**

\[ P(Y|X_1, \ldots, X_n) \]

**Discriminative**

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

**Generative**

\[ P(X_1, \ldots, X_n|Y), P(Y) \]
Generative vs discriminative models
NB decision rule

\( f : X \rightarrow Y \)
Most probable value of \( f(x) = y \):
NB decision rule

\[ f : X \rightarrow Y \]

Most probable value of \( f(x) = y \):

\[
\begin{pmatrix}
\mathbf{x}_1 \\
\mathbf{x}_2 \\
\vdots \\
\mathbf{x}_n \\
\end{pmatrix}
, \quad
\begin{pmatrix}
\mathbf{y}_1 \\
\mathbf{y}_2 \\
\vdots \\
\mathbf{y}_N \\
\end{pmatrix}
\]

\[ \rightarrow \text{crossover dribble} \]
\[ \rightarrow \text{crossover dribble} \]
\[ \rightarrow \text{shoot} \]
NB decision rule

\[ f : X \rightarrow Y \]

Most probable value of \( f(x) = y \):

\[
\begin{pmatrix}
  x_1^1 & \cdots & x_n^1 \\
  x_1^2 & \cdots & x_n^2 \\
  \vdots & \ddots & \vdots \\
  x_1^N & \cdots & x_n^N \\
\end{pmatrix},
\begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_N \\
\end{pmatrix} \rightarrow f(x_1^1, \ldots, x_n^1)
\]
Decision rule

\[ f : X \rightarrow Y \]

Most probable value of \( f(x) = y \):

\[ Y_{\text{predict}}^{\text{MLE}} = \]

\[ Y_{\text{predict}}^{\text{MAP}} = \]

\[ Y_{\text{predict}}^{\text{NB}} = \]
Decision rule

\[ f : X \to Y \]

Most probable value of \( f(x) = y \):

\[
Y_{\text{predict}}^{\text{MLE}} = \arg \max_{y_j \in Y} P(X = x_1, \ldots, X = x_n | Y = y_j)
\]

\[
Y_{\text{predict}}^{\text{MAP}} = \arg \max_{y_j \in Y} P(Y = y_j | X = x_1, \ldots, X = x_n)
\]

\[
Y_{\text{predict}}^{\text{NB}} = \arg \max_{y_j \in Y} \prod_{i}^{n} P(X = x_i | Y = y_j) P(Y = y_j)
\]
Decision rule

\[ f : X \rightarrow Y \]

Most probable value of \( f(x) = y \):

- Bayes optimal

\[
Y_{\text{ML}}^{\text{predict}} = \arg \max_{y_j \in Y} P(X = x_1, \ldots, X = x_n | Y = y_j)
\]

\[
Y_{\text{MAP}}^{\text{predict}} = \arg \max_{y_j \in Y} P(Y = y_j | X = x_1, \ldots, X = x_n)
\]

\[
Y_{\text{NB}}^{\text{predict}} = \arg \max_{y_j \in Y} \prod_{i}^{n} P(X = x_i | Y = y_j) P(Y = y_j)
\]
NB decision rule

\[ Y_{NB}^{predict} = \arg\max_{y_j \in Y} \prod_{i}^{n} P(X = x_i | Y = y_j) P(Y = y_j) \]

\[ P(X = x_i | Y = y_j) = \text{you know this from class} \]

\[ P(Y = y_j) = \text{you know this from class} \]

Matlab structures...
% example for NB - how to count
% number of data points
num_data_points = 9;
% number of possible values for the features
x_num_outcomes = 4;
% feature dimensionality
x_dim = 3;
% number of possible outcomes
y_num_outcomes = 2;

Y = [0 0 0 1 1 1 0 0 0]';
X = [ 4 3 4 3 4 3 4 3 4;
     1 4 3 1 2 1 2 3 1;
     3 1 2 4 1 4 1 3 2;
  ]';

fprintf('Data:
');
[X,Y]

% class conditional probability table
%(each_feature, y_outcomes, x_outcomes)
table = zeros(x_dim, y_num_outcomes, x_num_outcomes);

%P(X_i == k, Y = j) = ?
% what is this:
X(Y ==1, 3)

% some code....
% ....
    i = 1; % dim of X
    j = 1; % values of Y
    k = 1; % values of X
    table(i, j, k) = (sum( X(Y==j, i) == k ) + something ) / (sum( Y == j ) + something' );
% more code....

% once you have your conditional probability table, how do you make a decision?
NB decision rule

\[ Y_{NB}^{predict} = \arg \max_{y_j \in Y} \prod_{i} P(X = x_i | Y = y_j) P(Y = y_j) \]

In the homework: \( Y = 0 \) or \( Y = 1 \)

Comparing:

\[ \prod_{i} P(X = x_i | Y = 1) P(Y = 1) \geq \prod_{i} P(X = x_i | Y = 0) P(Y = 0) \]
$Y_{NB}^{\text{predict}} = \arg \max_{y_j \in Y} \prod_{i}^{n} P(X = x_i | Y = y_j) P(Y = y_j)$

Problem?

$0.2 \times 0.3 = \text{alright}$

$0.2 \times 0.3 \times 0.8 \times 0.1 \times 0.02 \times 0.7 \ldots = \text{trouble}$

Hint: “argmax” a monotonic function of the decision rule
NB: Gaussian inputs vs discrete inputs

Gaussian inputs: in class

Discrete inputs: in homework*

* See suggested reading for Gaussian NB
NB: performance

\[ f(X_1, X_2) \]

\[ Y \]

prediction \hspace{2cm} truth

Classification performance:

\[ \frac{\sum I(f(X_1, X_2) = Y)}{|Y|} \]

not the same as error rate...
NB: error rate

\[ f(X_1, X_2) \quad Y \]

prediction \quad truth

\[ f(X_1, X_2) \neq Y \]

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

\[ \sum_{X_1, X_2, Y} \mathbf{1}(f(X_1, X_2) \neq Y) P(X_1, X_2, Y) \]
Logistic regression

\[ P(Y = 1 | X, w) = g(w_0 + \sum_i w_i x_i) \]

\[ g(w_0 + \sum_i w_i x_i) = \frac{1}{1 + e^{w_0 + \sum_i w_i x_i}} \]

\[ g(w_0 + \sum_i w_i x_i) = g(w_0 + \sum_{i=1}^{n} w_i x_i) = g(\sum_{i=0}^{n} w_i x_i) \]

\[ x_0 = 1 \]
LR: learning the weights

MLE!

Log-likelihood: $lnP(D_Y \mid D_X, w)$

Concave: $lnP(D_Y \mid D_X, w) = \sum_j lnP(y^j \mid x^j, w)$

Gradient ascent: wikipedia or class on Monday
$ln P(D_y | D_x, w) = \sum_j ln P(y^j | x^j, w)$

$w \leftarrow w + \epsilon \frac{\delta}{\delta w} \sum_j ln P(y^i | x^i, w)$

For all features i:

$w_i^{t+1} \leftarrow w_i^t + \epsilon \sum_j x_i^j [y^j - P(Y^j = 1 | x^j, w^t)]$

Until... estimate doesn’t change “much”
Classification rule: see class notes
Information theory

Quick intro, this will be covered in class
Information theory

Entropy:  \( H(Y) = - \sum_{i=1}^{k} P(Y = y_i) \log_2 P(Y = y_i) \)

On average, smallest number of bits needed to transmit values drawn from Y’s distribution

Information gain:

\[ IG(Y|X) = H(Y) - H(Y|X) \]

How much better can we do if we share knowledge about X?
Information theory

Conditional entropy

\[
H(Y|X) = - \sum_{j=1}^{v} P(X = x_j) \sum_{i=1}^{k} P(Y = y_i | X = x_j) \log_2 P(Y = y_i | X = x_j)
\]
Decision trees

ID3

check wikipedia
Andrew Moore’s lectures:
http://www.autonlab.org/tutorials/

Carlos Guestrin’s lectures:
http://select.cs.cmu.edu/class/10701-F09/schedule.html

Tommi Jaakola lectures

Videos: CMU graphics lab
Homework questions?