Leren — Homework 4

Chapter 9-10, Alpaydin

Tim Stolp (11848782)

Deadline: 23:59 December 9th, 2018

This is the sixth week's assignment for Leren. This assignment covers chapters 9 & 10 of Alpaydin. Please take note of the following:

- You are expected to hand in your solutions in LaTeX;
- This problem set is an individual assignment;
- The deadline for this assignment is Sunday, December 9th, 2018 at 23:59.

1 Chapter 9: Decision Trees

Consider the training dataset given below. Based on this data, we would like to determine whether an example student is likely to succeed in an exam (represented by class variable Y). We make this decision solely based on two properties: whether the student is *hard-working* (represented by variable X_1) and whether the student *attends* all classes (represented by variable X_2). The '+' and '−' signs represent positive and negative classes, respectively.

Hard-working (X_1)	Attendance (X_2)	Success (Y)
Yes	Yes	+
Yes	No	−
Yes	No	+
Yes	Yes	+
No	Yes	−

(a) Is it possible to find a decision tree with 100% accuracy on this training set? If your answer is yes, draw the decision tree. If your answer is no, explain why.

It is impossible to get a decision tree with 100% accuracy. You will get stuck when you go down the tree with $x_1 = yes$ and $x_2 = no$. Here you will be left with a 50/50 chance of either success or no success.

(b) Compute the sample impurity $\mathcal{I}(Y)$ of the variable Y (success) using the entropy measure with logarithmic base 2.

$Entropy = -p * log_2(p) - (1-p) * log_2(1-p)$
$p = \frac{3}{5}$
$\mathcal{I}(+,-) = -\frac{3}{5} * log_2(\frac{3}{5}) - (1-\frac{3}{5}) * log_2(1-\frac{3}{5})$
$\mathcal{I}(+,-) = 0.442 + 0.529 = 0.971$

(c) Using entropy as a measure of the impurity in a collection of training examples, we define a measure of the effectiveness of an attribute in classifying the training data: The measure is called information gain and is the expected reduction in entropy caused by partitioning the examples according to this attribute.

For this case, we can express the information gain $IG(X_1)$ based on X_1 by computing the entropy of Y after a split on X_1, given by $\beth(Y|X_1)$:

$$IG(X_1) = \beth(Y) - \beth(Y|X_1)$$

The same thing for X_2:

$$IG(X_2) = \beth(Y) - \beth(Y|X_2)$$

What are the information gains $IG(X_1)$ and $IG(X_2)$ for this sample of training data?

$IG(X_1) = \beth(Y) - \beth(Y|X_1)$
$\beth(Y|X_1) = P(Yes)\beth(Yes(+,-)) + P(No)\beth(No(+,-))$
$= \frac{4}{5} * (-\frac{3}{4} * log_2(\frac{3}{4}) - (1-\frac{3}{4}) * log_2(1-\frac{3}{4})) + \frac{1}{5} * (-1 * log_2(1) - (1-1) * log_2(1-1))$
$= \frac{4}{5} * 0.811 + \frac{1}{5} * 0$
≈ 0.649
$IG(X_1) = 0.971 - 0.649 = 0.322$

$IG(X_2) = \beth(Y) - \beth(Y|X_2)$
$\beth(Y|X_2) = P(Yes)\beth(Yes(+,-)) + P(No)\beth(No(+,-))$
$= \frac{3}{5} * (-\frac{2}{3} * log_2(\frac{2}{3}) - (1-\frac{2}{3}) * log_2(1-\frac{2}{3})) + \frac{2}{5} * (-\frac{1}{2} * log_2(\frac{1}{2}) - (1-\frac{1}{2}) * log_2(1-\frac{1}{2}))$
$= \frac{3}{5} * 0.918 + \frac{2}{5} * 1$
≈ 0.951
$IG(X_2) = 0.971 - 0.951 = 0.020$

2 Chapter 10: Linear Discrimination

2.1 Logistic Regression

Consider a binary classification ($K = 2$ classes) model where it is given that

$$\log \frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} = \mathbf{w}^T\mathbf{x} + w_0\,, \tag{1}$$

i.e. the *log odds* of $P(\mathcal{C}_1|\mathbf{x})$ is linear. This is for example the case for two normal classes sharing a common covariance: $P(\mathbf{x}|\mathcal{C}_1) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $P(\mathbf{x}|\mathcal{C}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, and is independent of the choice for the class priors $P(\mathcal{C}_1)$ and $P(\mathcal{C}_2)$. Decision theory tells us that we should pick $\mathcal{C}_1$ if $P(\mathcal{C}_1|\mathbf{x}) > P(\mathcal{C}_2|\mathbf{x})$, assuming a 0/1 loss.

(a) Simplify the left-hand side of Eq. (1) to express it only in terms of $P(\mathcal{C}_1|\mathbf{x})$ and $P(\mathcal{C}_2|\mathbf{x})$.

$P(\mathcal{C}_1|\mathbf{x}) \propto P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)$ (Bayes)
$P(\mathcal{C}_2|\mathbf{x}) \propto P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)$ (Bayes)
$\log \frac{P(\mathcal{C}_1|\mathbf{x})}{P(\mathcal{C}_2|\mathbf{x})} = \mathbf{w}^T\mathbf{x} + w_0$

(b) Solve for $P(\mathcal{C}_1|\mathbf{x})$. Hint, in the expression one can re-write $P(\mathcal{C}_2|\mathbf{x})$ as: $P(\mathcal{C}_2|\mathbf{x}) = 1 - P(\mathcal{C}_1|\mathbf{x})$.

$\log(\frac{P(\mathcal{C}_1|\mathbf{x})}{P(\mathcal{C}_2|\mathbf{x})}) = \log(\frac{P(\mathcal{C}_1|\mathbf{x})}{1-P(\mathcal{C}_1|\mathbf{x})}) = \log(\frac{1}{\frac{1}{P(\mathcal{C}_1|\mathbf{x})}-1}) = -\log(\frac{1}{P(\mathcal{C}_1|\mathbf{x})} - 1)$
$-\log(\frac{1}{P(\mathcal{C}_1|\mathbf{x})} - 1) = \mathbf{w}^T\mathbf{x} + w_0$
$\log(\frac{1}{P(\mathcal{C}_1|\mathbf{x})} - 1) = -(\mathbf{w}^T\mathbf{x} + w_0)$
$\frac{1}{P(\mathcal{C}_1|\mathbf{x})} - 1 = \exp(-(\mathbf{w}^T\mathbf{x} + w_0))$
$\frac{1}{P(\mathcal{C}_1|\mathbf{x})} = \exp(-(\mathbf{w}^T\mathbf{x} + w_0)) + 1$
$P(\mathcal{C}_1|\mathbf{x}) = \frac{1}{\exp(-(\mathbf{w}^T\mathbf{x}+w_0))+1}$

(c) You can rewrite your result from (b) as $P(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{a})$, where $\sigma(\mathbf{a})$ is the sigmoid function, defined as $\sigma(\mathbf{a}) = \frac{1}{1+\exp(-\mathbf{a})}$. Provide an expression for $\mathbf{a}$ in terms of $\mathbf{x}$, $\mathbf{w}$ and w_0.

$\mathbf{a} = \mathbf{w}^T\mathbf{x} + w_0$

(d) What is the decision boundary in terms of $P(\mathcal{C}_1|\mathbf{x})$? Further provide the decision boundary in terms of the log odds of $P(\mathcal{C}_1|\mathbf{x})$, i.e. in terms of $\mathbf{w}^T\mathbf{x} + w_0$. Provide your solution in the form: classify $\mathbf{x}$ as $\mathcal{C}_1$ if $P(\mathcal{C}_1|\mathbf{x}) > a$ (or $\mathbf{w}^T\mathbf{x} + w_0 > b$), where a (or b) is a number.

The decision boundary in terms of $P(\mathcal{C}_1|\mathbf{x})$:
classify $\mathbf{x}$ as $\mathcal{C}_1$ if $P(\mathcal{C}_1|\mathbf{x}) > 1 - P(\mathcal{C}_1|\mathbf{x})$

The decision boundary in terms of the log odds of $P(\mathcal{C}_1|\mathbf{x})$:
classify $\mathbf{x}$ as $\mathcal{C}_1$ if $\mathbf{w}^T\mathbf{x} + w_0 > 0$

2.2 Derivative of Softmax

Show that the derivate of the softmax, $y_i = \exp(a_i)/\sum_j \exp(a_j)$ is $\partial y_i/\partial a_j = y_i(\delta_{ij} - y_j)$, where δ_{ij} is 1 if $i = j$ and 0 otherwise.

$y_i = \frac{g(x)}{h(x)}$ with $g(x) = \exp(a_i)$ and $h(x) = \sum_j \exp(a_j)$
$\partial y_i/\partial a_j = \frac{g'(x)h(x) - h'(x)g(x)}{h(x)^2}$
$h'(x) = \exp(a_j)$
$g'(x) = \exp(a_i)$ if $i = j$ else $g'(x) = 0$

if $i = j$:
$\frac{\exp(a_i)\sum_j \exp(a_j) - \exp(a_j)\exp(a_i)}{(\sum_j \exp(a_j))^2} = \frac{\exp(a_i)}{\sum_j \exp(a_j)} \frac{\sum_j \exp(a_j) - \exp(a_j)}{\sum_j \exp(a_j)} = \frac{\exp(a_i)}{\sum_j \exp(a_i)} (\frac{\sum_j \exp(a_j)}{\sum_j \exp(a_j)} - \frac{\exp(a_j)}{\sum_j \exp(a_j)})$
$= y_i(1 - y_j)$

if $i \neq j$:
$\frac{0 - \exp(a_j)\exp(a_i)}{(\sum_j \exp(a_j))^2} = -\frac{\exp(a_i)}{\sum_j \exp(a_j)} \frac{\exp(a_j)}{\sum_j \exp(a_j)}$
$= -y_i y_j$

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Gives $\partial y_i/\partial a_j = y_i(\delta_{ij} - y_j)$