Lecture 2: Inductive learning

Helen Yannakoudakis and Oana Cocarascu

Department of Informatics
King's College London

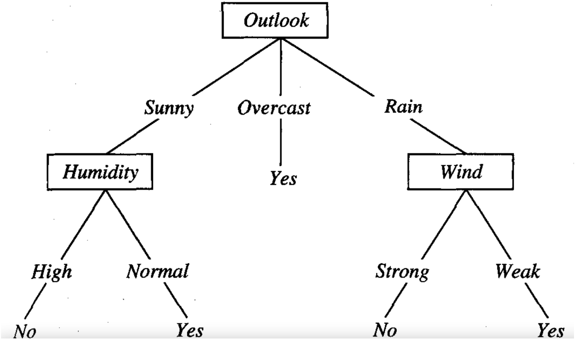
January 24, 2022

Lecture 2 Q&A

Can you go over an example of Rule post-pruning? / Could you explain rule post-pruning in more detail? More specifically, I don't understand how to sort the rules by accuracy. (1) How do you measure the accuracy of a rule? How are ties broken? (2) Which feature tests to remove from the rules?

(Mitchell Chapter 3 (3.7.1.2))

Lecture 2 Q&A



Outlook=sunny AND Humidity=high → PlayTennis=No

Outlook=sunny → PlayTennis=No

Humidity=high → PlayTennis=No

Lecture 2 Q&A

Can you please explain univariate and multivariate linear regression with examples? / Can you go through an example of multivariate linear regression?

Lecture 2 Q&A

- Univariate equation is of the form:

$$h_w(x) = w_1 x + w_0$$

Example: predicting house prices by floor area.

- In multivariate, we have more variables/features:
- So $h_w(x_j)$ is now the weighted sum of the variable values:

$$h_w(x) = \sum_{i=0}^{i=n} w_i x_i = w^\intercal x$$

We estimate w from data.

Lecture 2 Q&A

Would you be able to explain the more sophisticated KNN equation? / Could you please give an example using the sophisticated formula version of the K nearest neighbour(slide 38 Lecture 1)?

(Murphy's book, chapter 16.1)

- Use the k nearest points to estimate the probability of class membership:

$$p(y = c|\mathbf{x}, \mathcal{D}, K) = \frac{1}{K} \sum_{i \in N_K(\mathbf{x}, \mathcal{D})} \mathbb{I}(y_i = c)$$

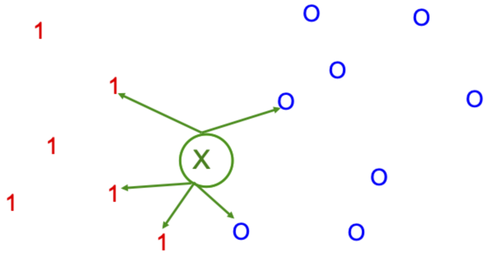
- $\mathbb{I}(e)$ is an indicator function such that:

$$\mathbb{I}(e) = \begin{cases} 1 & \textit{if e is true} \\ 0 & \textit{if e is false} \end{cases}$$

- Counts how many members of each class are in the k nearest set.

Lecture 2 Q&A

knn example for $K = 5$, where x is our test point, and the nearest neighbours of x have labels $\{1, 1, 1, 0, 0\}$:



3 of the 5 nearest neighbours have label 1;
2 of the 5 have label 0.

- $p(y = 1|\mathbf{x}, \mathcal{D}, K) = 3/5 = 0.6$
- $p(y = 0|\mathbf{x}, \mathcal{D}, K) = 2/5 = 0.4$

Lecture 2 Q&A

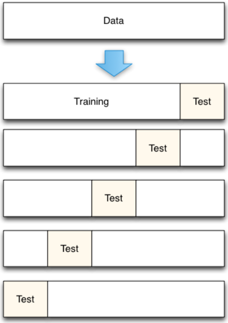
Could you please explain the difference between parametric and non-parametric techniques? Could you please give some more examples of both types?
/ Would k-means be considered parametric, because by setting k to some number you are assuming something about the structure of the data (how many clusters there are)?

Lecture 2 Q&A

Can you explain K-fold cross-validation again please?

Lecture 2 Q&A

- 5-fold cross-validation:
- Split data into $k = 5$ equal and unique subsets (folds).
- Train your model on $k - 1 = 4$ sets (combined together) and test on the remainder 1 fold only.
- Repeat above until you test on each fold only once (total $k = 5$ times)
- Compute misclassification error averaged over all $k = 5$ test folds.

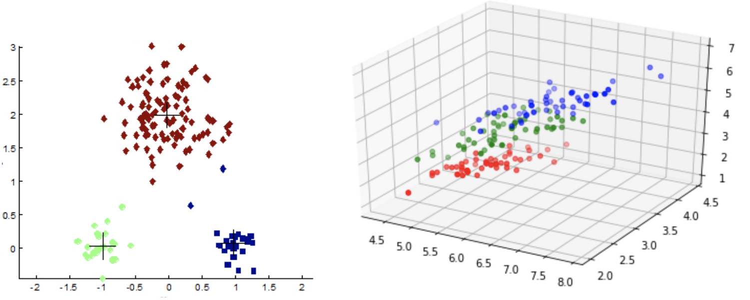


Can we discuss a full example of the C4.5 algorithm?

Lecture 2 Q&A

Are we expected to know 3 dimensional clustering?If so, could you please go over an example?

(Murphy's book chapter 21).



(a) (b)

Figure: Clustering examples.

Could you explain Batch gradient descent and stochastic gradient descent with an example for context?

- Batch: a each step we consider all the training examples, and update the weights using:

$$
\begin{aligned}
w_0 &\leftarrow w_0 + \alpha \sum_j (y_j - h_w(x_j)) \\
w_1 &\leftarrow w_1 + \alpha \sum_j (y_j - h_w(x_j))x_j
\end{aligned}
$$

- Stochastic: we can just do:

$$
\begin{aligned}
w_0 &\leftarrow w_0 + \alpha (y - h_w(x)) \\
w_1 &\leftarrow w_1 + \alpha (y - h_w(x))x
\end{aligned}
$$

For each of the N examples in turn.

Lecture 2 Q&A

Please can you explain the logistic regression update rule and cross-entropy loss? I've seen the extra slides but I am still very much lost!

Lecture 2 Q&A

Explain the (derivative term for the) logistic regression update rule.

- Use the logistic function as a threshold: $h_{\mathbf{w}}(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}}$
- Update rule: $w_i \leftarrow w_i - \alpha \frac{\partial}{\partial w_i} Loss(\mathbf{w})$
- Equations for a single training example:

$$\frac{\partial}{\partial w_i}(y - h_{\mathbf{w}}(\mathbf{x}))^2$$

$$2\,(y - h_{\mathbf{w}}(\mathbf{x})) \times \frac{\partial}{\partial w_i}(y - h_{\mathbf{w}}(\mathbf{x}))$$

$$-2\,(y - h_{\mathbf{w}}(\mathbf{x})) \times g'(\mathbf{w}\cdot\mathbf{x}) \times x_i$$

where g' is the derivative of the logistic function.

- Update rule becomes:

$$w_i \leftarrow w_i + \alpha(y - h_w(x)).h_w(x)(1 - h_w(x)).x_i$$

Lecture 2 Q&A

Could you explain the cross-entropy formula?

Logistic regression commonly uses **cross-entropy loss** function.

$p(y|x) = \hat{y}^y \, (1 - \hat{y})^{1-y}$ where $y \in \{0, 1\}$ and $\hat{y} = h(x)$

Rewrite as:

$log\, p(y|x) = y\, log\, \hat{y} + (1 - y)\, log\, (1 - \hat{y})$

Flip sign to turn into loss:

$-log\, p(y|x) = -[y\, log\, \hat{y} + (1 - y)\, log\, (1 - \hat{y})]$