Tutoring Session 2

Decision Trees and k-Nearest Neighbors

1 Building Decision Trees

Problem 1: Build a decision tree for the dataset $\mathcal{D}$[1] below.

Day	Outlook	Temperature	Humidity	Wind	Play Tennis?
D1	sunny	hot	high	weak	No
D2	sunny	hot	high	strong	No
D3	overcast	hot	high	weak	Yes
D4	rain	mild	high	weak	Yes
D5	rain	cool	normal	weak	Yes
D6	rain	cool	normal	strong	No
D7	overcast	cool	normal	strong	Yes
D8	sunny	mild	high	weak	No
D9	sunny	cool	normal	weak	Yes
D10	rain	mild	normal	weak	Yes
D11	sunny	mild	normal	strong	Yes
D12	overcast	mild	high	strong	Yes
D13	overcast	hot	normal	weak	Yes
D14	rain	mild	high	strong	No

Use the *ID3* algorithm. In contrast to CART, ID3 allows for multiway splits and exhausts all possible values for a feature when that feature is chosen for a split. The criterion that determines the best split is information gain, which in turn is based on entropy.

Problem 2: We consider decision trees for a two-class classification problem with classes 0 and 1. Let $\Phi(p, q)$ be a strictly concave function defined on $0 \leq p, q \leq 1$ such that

- $\Phi(1, 0) = \Phi(0, 1)$ is minimal;
- $\Phi(\frac{1}{2}, \frac{1}{2})$ is maximal.

Then, for $i(t) = \Phi(p(c = 0 \mid t), p(c = 1 \mid t))$, $\Delta i(s, t) = i(t) - p_R\, i(t_R) - p_L\, i(t_L)$ and any split s, show that

$$\Delta i(s, t) \geq 0,$$

with equality if and only if $p(c = i \mid t) = p(c = i \mid t_L) = p(c = i \mid t_R)$ for both $i = 0, 1$[2].

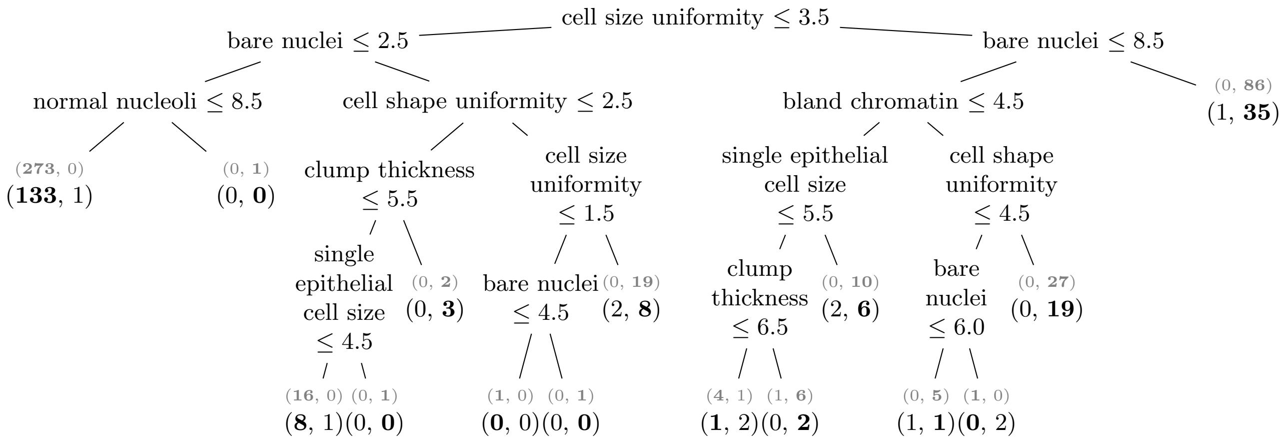
Hint: Strict concavity for $\Phi(p, q)$ means that for p_1, q_1, p_2, q_2 and $\alpha \in [0, 1]$

$$\Phi((1-\alpha)p_1 + \alpha p_2, (1-\alpha)q_1 + \alpha q_2) > (1-\alpha)\Phi(p_1, q_1) + \alpha\Phi(p_2, q_2)$$

[1]from T. Mitchell. 1997. Machine Learning. McGraw Hill
[2]adapted from G. Louppe. 2014. Understanding Random Forests. PhD Thesis

2 Pruning

You trained a tree on 455 samples of the popular Wisconsin breast cancer dataset to a maximum depth of 5 and noted the number of samples of each class in the leaves as tuples $(n_{\text{benign}}, n_{\text{malignant}})$ in gray font, indicating the class label for each leaf in bold. You realize that many leaves contain only a small number of samples and decide that you want to prune the tree. Luckily, you held out 228 samples during training so that you can now perform reduced error pruning on the final tree. You note down the number of validation samples in every leaf in black font and start pruning.



Problem 3: Which nodes do you prune? What are the error rates on both your training and validation set before and after pruning?

Problem 4: Now assume that the consequences (cost!) of misclassifying malignant as benign (type II error) are 10 times as high as classifying benign as malignant (type I error). Do you still prefer the pruned version of your tree?

3 Random Forests

In this exercise we will investigate the effect of two main parameters in Random Forests, namely the number of trees and the number of features randomly chosen at each node.

Problem 5: Suppose a random forest classifier discriminates between only two different classes. Assuming that the outputs of all trees are independent and have the same individual error rate ϵ, write down the probability of a random forest making the wrong prediction given the (odd!) number of trees n and the error rate of a single tree ϵ.

Problem 6: In practice, the limited number of training samples does not allow us to build a large number of independent trees. Consider a Random Forest with $n = 35$, where sets of five trees are mutually independent but where each of the trees within a set has the same output. How does this affect the accuracy of the random forest in contrast to a Random Forest with all independent trees?

Problem 7: Every tree of a Random Forest is built using a bootstrap sample of n samples chosen drawn from n available samples with replacement. Show that the probability that a certain sample s is used to build a certain tree T is $p(s \in T) \approx .632$ for large n.

Problem 8: If you have $n = 10$ samples, how many different bootstrap samples are possible?

Problem 9: Apart from taking bootstrap samples, variance between trees is introduced by randomly sampling only $d < D$ of the total number of D features at any node to determine the best split. Which effect does the parameter d have on the error rates?

4 Probabilistic k-NN

Assume that you have two classes. Let N_0 be the number of samples in class 0, N_1 the number of samples in class 1 (this implies that $p(c = 0) = \frac{N_0}{N_0+N_1}$ and $p(c = 1) = \frac{N_1}{N_0+N_1}$). Let $\boldsymbol{x}^*$ be a point that you want to classify.

Problem 10: Consider the ratio $\frac{p(c=0|\boldsymbol{x}^*)}{p(c=1|\boldsymbol{x}^*)}$. Show that for small σ^2 the following approximation holds:

$$\frac{p(c = 0 \mid \boldsymbol{x}^*)}{p(c = 1 \mid \boldsymbol{x}^*)} \approx \frac{\exp\left((-\|\boldsymbol{x}^* - \boldsymbol{x}_0\|^2)/(2\sigma^2)\right)}{\exp\left((-\|\boldsymbol{x}^* - \boldsymbol{x}_1\|^2)/(2\sigma^2)\right)}$$

Hint: You may assume that if σ^2 is very small, then the closest data points for each class (denoted by $\boldsymbol{x}_0$ for class 0 and $\boldsymbol{x}_1$ for class 1) will dominate the sum over the exponentials.

Problem 11: Show that in the limit case ($\sigma \to 0$) this approximation classifies $\boldsymbol{x}^*$ as class 0 if it is closer to $\boldsymbol{x}_0$ than to $\boldsymbol{x}_1$.

Problem 12: How does σ relate to k?

5 Neighbourhod Component Analysis

Problem 13: Calculate the gradient of the NCA objective as given in the slides. The following matrix identity[3] is helpful:

$$\frac{\partial tr(A^T AB)}{\partial A} = A(B + B^T)$$

tr is the *trace* of a matrix. You may also want to use the shorthand $x_{ij} = (x_i - x_j)$.

[3] for more of these things: K. B. Petersen and M. S. Pedersen. 2008. The Matrix Cookbook. Technical Report.