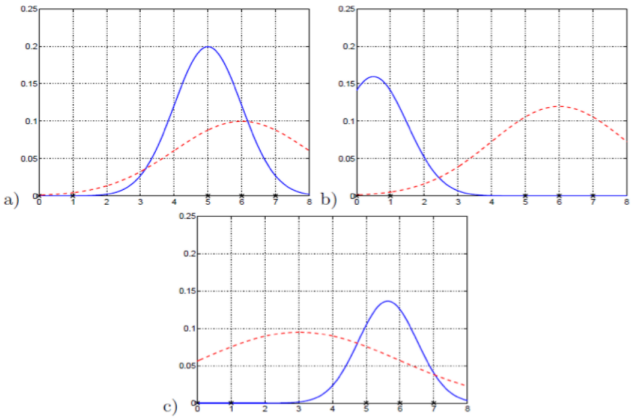
Theoretical task 2: data analyst routines

Due Monday, January 26.

1. Alice and Bob argue whether adding extra feature may decrease classification accuracy on the training set or not. Alice states that it cannot, since machine learning algorithm has complete information about training set and it will always take advantage of new information. Bob feels skeptical about that, especially if new information is not relevant for classification target. Who is right? If Alice - please, give your reasoning, if Bob - give a simple illustrative example.
2. Bob says that it is always better to standardize features by rescaling either making their mean equal to zero and variance equal to unity, or making them lie inside [0,1] interval. Alice argues that some algorithms perform inner rescaling compensating input feature scale, so that independently of input scale their output will be the same. Which of the following algorithms have such property? Why?
   - QDA
   - K-NN with L_∞ metric
   - K-NN with Mahalonobis metric
   - decision tree
3. Alice says that decision tree global optimality may be under question due to the fact that at each node construction only limited one-step ahead optimization is performed. Bob argues that successive application of 1-step optimizations iteratively leads to globally optimal solution (measured in terms of finally achieved impurity function value). Who is right? Please, provide your proof or simple illustrative counterexample.
4. Alice performed EM-algorithm to fit sample, consisting of 4 points with a mixture of 2 Gaussian distributions. Bob asked Alice to show the graphs, illustrating initial state of EM and the first step of EM. Alice said that these 2 illustrations together with one irrelevant graph lie on the table in her room. When Bob came there, he found that these three illustrations were not labelled. How can you logically deduce which figure stands for what? Why?



5. Bob asked Alice to perform unsupervised clustering using K-means algorithm for the objects $x_1, x_2, ...x_N$. The complication was that only pairwise distance information $d(x_i, x_j)$ was available, not feature vectors themselves. Alice found that she could not calculate cluster centers without feature representations. Bob suggested that she could consider only training objects as possible cluster centers. How could you reasonably adapt K-means algorithm to this situation using Bob's suggestion? Will time complexity of the new algorithm change? Why?