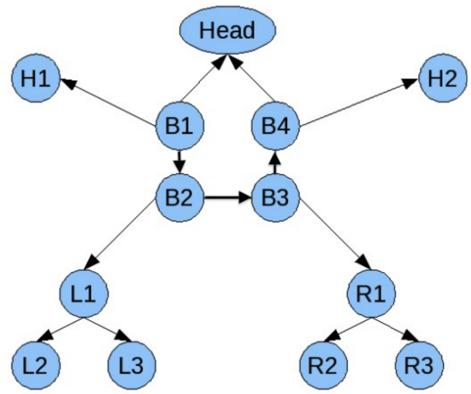
Final exam

Special Topics in Advanced Machine Learning
Spring 2017
Instructor: Anna Choromanska

Problem 1 (100 points)

Eve is looking for WallE using her cameras but can't find WallE. Eve has small circuits for performing the junction-tree algorithm. Help her out by designing a junction-tree from the graph below which Eve has in her mind for WallE.



Problem 2 (40 points)

A kernel is an efficient way to write out an inner product between two feature vectors computed from a pair of input vectors as follows:

$$K(x, y) = \phi(x)^\top \phi(y).$$

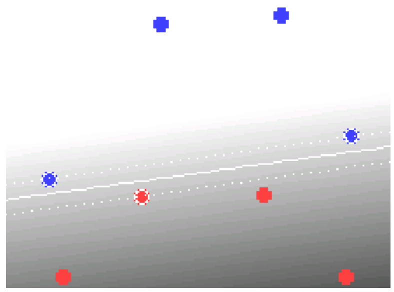
Assume that both inputs are 2-dimensional and write out the explicit mapping ϕ that mimics the kernel value for a 3$^{\text{rd}}$-order polynomial kernel as follows:

$$K(x, y) = (x^\top y + 1)^3.$$

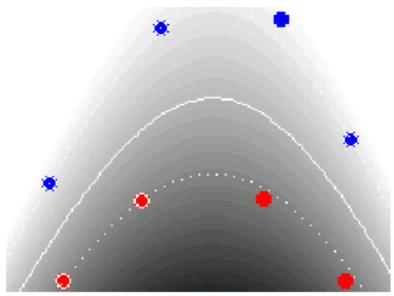
Problem 3 (30 points)

Assume we have trained 3 separable support vector machines on the 2D data (the axes go from -1 to 1 in both horizontal and vertical direction) using 3 different kernels as follows:

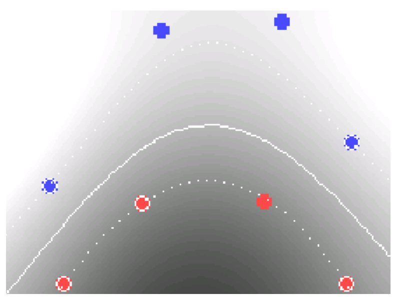
a) (10 points) a linear kernel (i.e. the standard linear SVM): $k(x_i, x_j) = x_i^\top x_j$



b) (10 points) a quadratic polynomial kernel: $k(x_i, x_j) = (x_i^\top x_j + 1)^2$



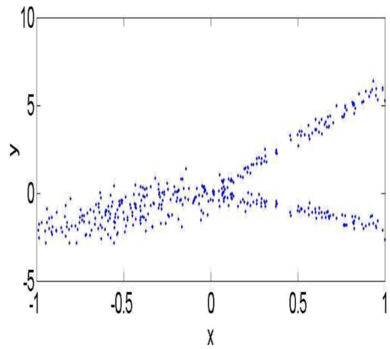
c) (10 points) an RBF kernel: $k(x_i, x_j) = \exp\left(-\frac{1}{2\sigma}\|x_i - x_j\|^2\right)$



Assume we now translate the data by adding a large constant value (i.e. 10) to the vertical coordinate of each data points, i.e. a point (x_1, x_2) becomes $(x_1, x_2 + 10)$. If we retrain the above SVMs on this new data, how does the resulting SVM boundary change relative to the data points? Explain why or why not it changes for all 3 cases (a), (b), and (c) and draw what happens to the resulting new boundaries when appropriate.

Problem 4 (50 points)

You are given the data set in the figure below which is fit with maximum likelihood via EM using a mixture of 3 Gaussians. Assume EM converged nicely to the optimal maximum likelihood solution. Draw the 3-Gaussian fit you would expect on top of the data below. (10 points)



EM thus gives us the following joint distribution:

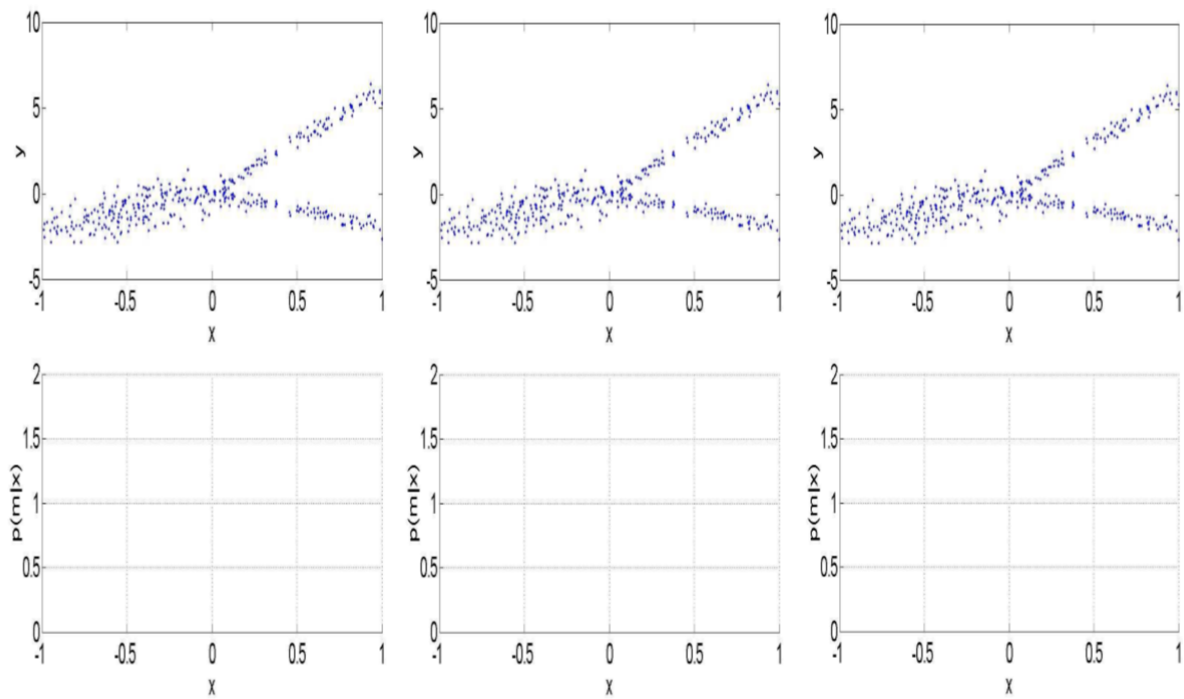
$$p(x, y) = \sum_m p(m, x, y) = \sum_{m=1}^{3} \alpha_m \mathcal{N}(x, y|\mu_m, \Sigma_m).$$

The mixture model is conditioned to form a mixture of experts conditional pdf:

$$p(y|x) = \sum_m p(m, y|x) = \sum_{m=1}^{3} p(m|x)p(y|m, x).$$

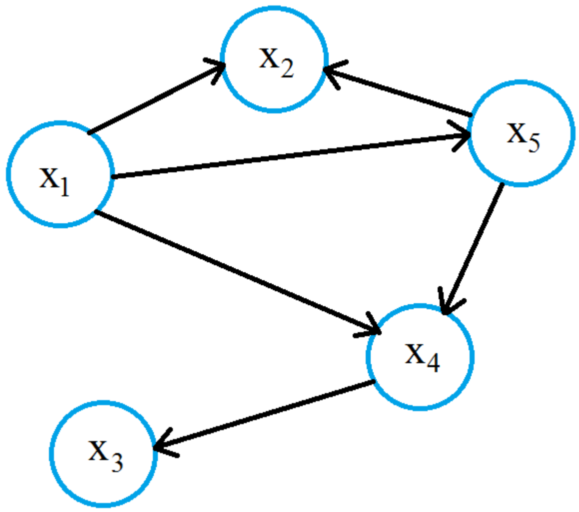
Assume the original Gaussians give rise to Gates $p(m|x)$ functions as above and the conditioned Gaussians give rise to the Experts $p(y|m, x)$. In the 3 figures below, draw three expert/gate combinations (i.e. $p(y|x, m)$ and $p(m|x)$) for $m = 1$, $m = 2$, and $m = 3$. The order ($m = 1, 2, 3$) of the experts/gates doesn't matter. Plot each expert as a contour plot of the conditional probability of y given m and x as x, y varies and plot the value of $p(m|x)$ for each gate as x varies. (30 points) Briefly explain your answer. (10 points)

$p(y|x, m = 1)$ & $p(m = 1|x)$ $p(y|x, m = 2)$ & $p(m = 2|x)$ $p(y|x, m = 3)$ & $p(m = 3|x)$



Problem 5 (110 points)

Consider the Bayesian network below with binary variables $x_1, x_2, \ldots, x_5$.



Write out the factorization of the probability distribution $p(x_1, ..., x_5)$ implied by this directed graph. (10 points) Then, using the Bayes ball algorithm, indicate for each statement below if it is True or False and justify your answers (100 points)

- x_2 and x_4 are independent.
- x_2 and x_4 are conditionally independent given x_1, x_3, and x_5.
- x_2 and x_4 are conditionally independent given x_1 and x_3.
- x_5 and x_3 are conditionally independent given x_4.
- x_5 and x_3 are conditionally independent given x_1, x_2, and x_4.
- x_1 and x_3 are conditionally independent given x_5.
- x_1 and x_3 are conditionally independent given x_2.
- x_2 and x_3 are independent.
- x_2 and x_3 are conditionally independent given x_5.
- x_2 and x_3 are conditionally independent given x_5 and x_4.

Problem 6 (110 points)

Show the convergence guarantee (along with all the derivations) of the following:

a) (40 points) Randomized iterative optimization algorithm that at iteration T, where $T = 1, 2, \ldots$, obtains parameter vector x_T and outputs the average of all previously obtained parameter vectors, i.e. it outputs $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} x_t$. Furthermore, you know that at any time t the following is satisfied:

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{1}{t},$$

where f is convex and has optimum at x^*. Name the convergence rate of this algorithm.

b) (40 points) Iterative optimization algorithm that at any iteration t, where $t = 1, 2, \ldots$, satisfies:

$$\|x_t - x^*\|_2 \leq \frac{1}{4}\alpha\|x_{t-1} - x^*\|_2,$$

where x^* is an optimum. What is the condition on α to ensure the convergence of this algorithm? Name the convergence rate of this algorithm for the plausible setting of α.

c) (30 points) What optimization algorithm achieves quadratic convergence rate? Provide the update of this algorithm and its computational complexity.

Problem 7 (30 points)

Show the first two iterations (after the initialization) of the k-means clustering algorithm (show centers and assignments of data points to clusters) for the following 2D data set: $(4, 2), (0, -1), (1, 4), (2, 8), (3, 5), (8, 8), (3, 3),$ $(10, 10), (20, 18)$, and $(12, 9)$. Assume the number of centers is equal to 2 and the centers are initialized to $(1, 1)$ and $(7, 8)$.

Problem 8 (30 points)

What is the VC dimension of the hypothesis space consisting of triangles in the 2D plane (justify your answer)? Points inside the triangle are classified as positive examples.