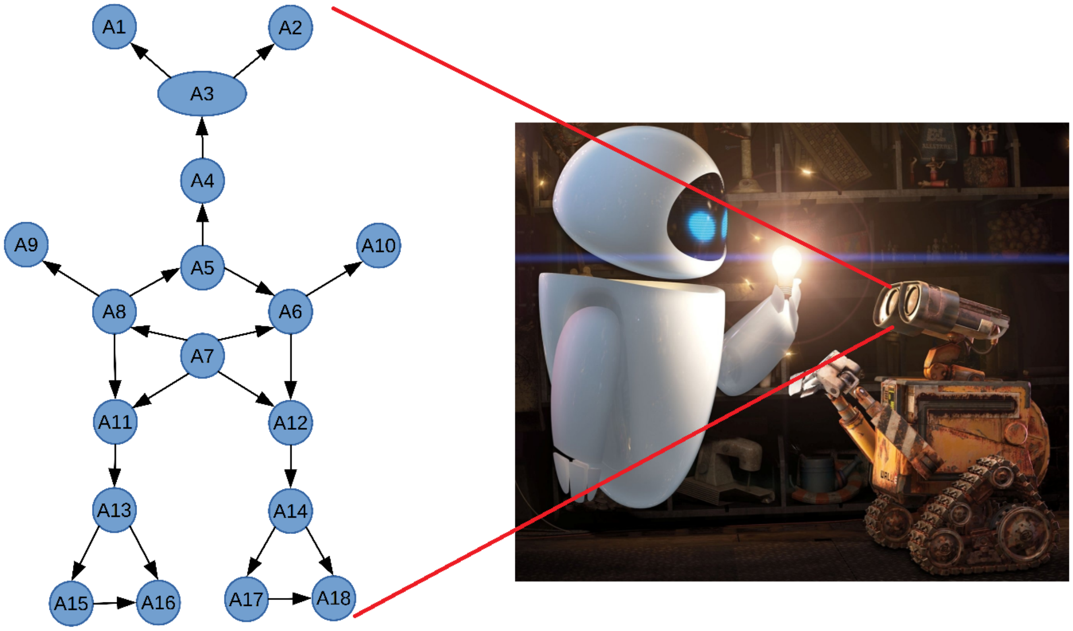
Final exam

Introduction to Machine Learning
Fall 2018
Instructor: Anna Choromanska

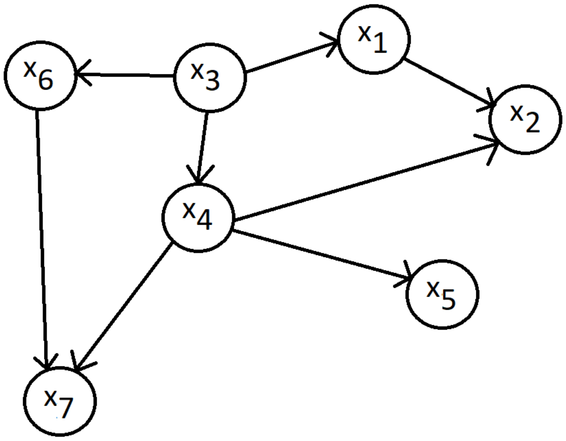
Problem 1 (90 points)

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



Problem 2 (60 points)

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



Write out the factorization of the probability distribution $p(x_1, ..., x_7)$ 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 (50 points)

- x_2 and x_6 are independent.
- x_2 and x_6 are conditionally independent given x_1, x_3, and x_5.
- x_1 and x_7 are conditionally independent given x_5.
- x_5 and x_3 are conditionally independent given x_1 and x_2.
- x_5 and x_6 are conditionally independent given x_1, x_2, and x_4.
- x_5 and x_6 are conditionally independent given x_4.
- x_5 and x_6 are conditionally independent given x_1.
- x_2 and x_6 are conditionally independent given x_3 and x_5.
- x_1 and x_7 are independent.
- x_1 and x_7 are conditionally independent given x_4.

Problem 3 (40 points)

Consider the fragment of the convolutional architecture given below:

- Input image: $1 \times x \times y$
- Convolutional layer: $\underbrace{1 \to 4}_{\text{number of input and output channels}}, \underbrace{5 \times 5}_{\text{filter size}}, \underbrace{2 \times 3}_{\text{stride}}$
- ReLU
- MaxPooling: $\underbrace{3 \times 3}_{\text{region size}}, \underbrace{3 \times 3}_{\text{stride}}$
- Convolutional layer: $4 \to 6, 4 \times 4, 2 \times 2$
- ReLU
- MaxPooling: $2 \times 2, 2 \times 2$
- Flattening (3D to 1D):
  $\underbrace{6 \times 12 \times 8}_{\text{number of feature maps } \times \text{ size of the feature map } (12 \times 8)} \to 576$

What is the size of the input (in other words what is x and y)?

Problem 4 (20 points)

Consider the following plot, where we fit the polynomial of order M ($f(x;w) = \sum_{j=0}^{M} w_j x^j$) to the dataset, where $w = [w_0 \quad w_1 \quad \ldots \quad w_M]^\top$ denotes the vector of model weights and the dataset is a collection of 2-dimensional points (x, y). The dataset is represented with the blue circles on the figure.

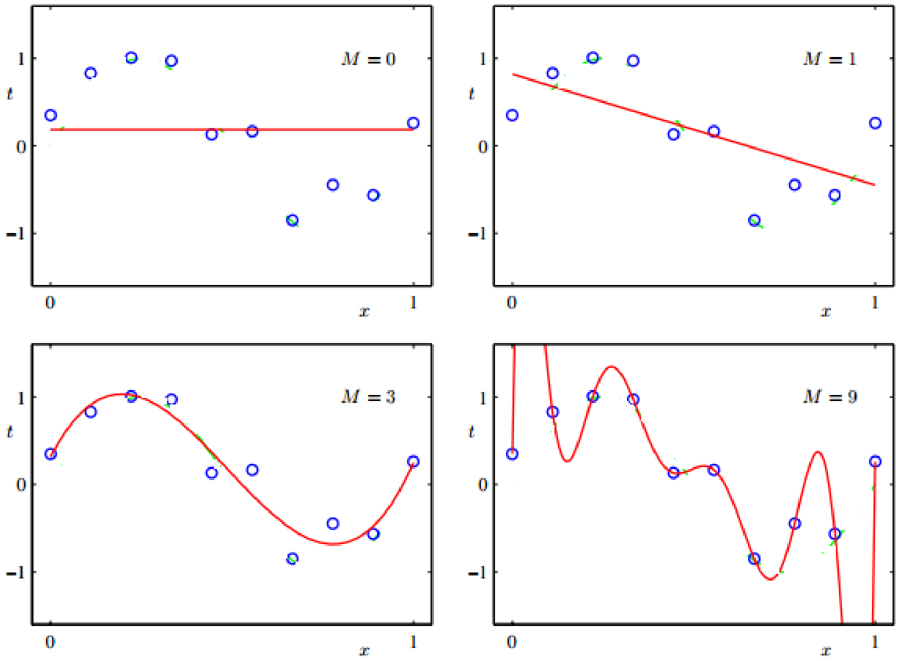


Figure 1: Plots of polynomials having various orders M, shown as red curves, fitted to the data set.

What is the reasonable choice of M and why? Which M correspond to overfitting and which to underfitting and why?

Consider any loss function that measures the discrepancy between the target values and the predictions of the model, e.g. squared loss which for a single data point is defined as $L(y_i, f(x_i, w)) = \frac{1}{2}(y_i - f(x_i, w))^2)$. Draw a typical behavior of the train and test loss for the optimal setting of model weights as a function of M, where recall that the train loss is the loss computed for a training dataset (the model was trained on this dataset) and the test loss is the loss computed for a test dataset (the model did not see this dataset during training). Indicate overfitting and underfitting regimes.

Problem 5 (100 points)

You are given the parameters of a 2-state HMM. You observed the input sequence AB (from a 2-symbol alphabet A or B). In other words, you observe two symbols from your finite state machine, A and then B. Using the junction tree algorithm, evaluate the likelihood of this data $p(y)$ given your HMM and its parameters. Also compute (for decoding) the individual marginals of the states after the evidence from this sequence is observed: $p(q_0|y)$ and $p(q_1|y)$. The parameters for the HMM are provided below. They are the initial state prior $p(q_0)$, the state transition matrix given by $p(q_t|q_{t-1})$, and the emission matrix $p(y_t|q_t)$, respectively.

$$\pi = p\left(q_0\right) = \begin{array}{cc} 1 & 2 \\ \multicolumn{2}{c}{\left[\begin{array}{cc} 1/4 & 3/4 \end{array}\right]} \end{array}$$

$$a^T = p\left(q_t \mid q_{t-1}\right) = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \left[\begin{array}{cc} 1/2 & 1/3 \\ 1/2 & 2/3 \end{array}\right] \end{array} \qquad \eta^T = p\left(y_t \mid q_t\right) = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \left[\begin{array}{cc} 1/4 & 1/8 \\ 3/4 & 7/8 \end{array}\right] \end{array}$$

Problem 6 (20 points)

Consider 2d family of classifiers given by axis-aligned squares. What is the VC dimension of this family?