UNIVERSITY OF SOUTHAMPTON COMP3008A

SEMESTER 2 EXAMINATION 2004/2005

MACHINE LEARNING

Duration: 120 mins

Answer ALL *questions from section A (20 marks)*
and ONE *question from section B (25 marks)*
and ONE *question from section C (25 marks).*

This examination is worth 70%. The coursework was worth 30%.

Calculators without text storage MAY be used.

 Number of pages 11

Section A

Question 1

(a) Compute the update rule $\boldsymbol{w}' = \boldsymbol{w} - \eta \boldsymbol{\nabla} E(\boldsymbol{w})$ where $E(\boldsymbol{w})$ consists of a squared error term and a regularisation term

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{k=1}^{P} \big(g(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_k) - y_k\big)^2 + \frac{\gamma}{2} \|\boldsymbol{w}\|^2.$$

Explain why the regularisation term is often referred to as a weight decay term.

Test ability to compute simple gradients. This was a typical calculation shown in the course.

$$\boldsymbol{\nabla} E(\boldsymbol{w}) = \sum_{k=1}^{P} \big(g(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_k) - y_k\big)\; g'(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_k)\, \boldsymbol{w} + \gamma \boldsymbol{w}$$

so that

$$\boldsymbol{w}' = (1 - \eta\,\gamma)\boldsymbol{w} - \eta \sum_{k=1}^{P} \big(g(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_k) - y_k\big)\; g'(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_k)\, \boldsymbol{w}$$

The regularisation causes the weights to be decreased (by a factor $1 - \eta\,\gamma$) at every learning step.

(5 marks)

(b) Classify the point $(8, 0.3)$ using a 3-nearest-neighbours algorithm where you have a training set of positive and negative examples
$\mathcal{D}^+ = \{(6, 0.8), (7, 0.9), (9, 0.8), (10, 0.9)\}$
$\mathcal{D}^- = \{(6, 0.3), (5, 0), (4, 0.4), (0, 0.2)\}$.
Normalise the data to the unit square and recompute the classification. Explain the importance of normalising data.

Tests understanding of scaling in the context of data. Let $\boldsymbol{x} = (8, 0.3)$ and $\boldsymbol{x}_i$ denote a vector in the data set. Prime denotes the normalised data. Then

$\mathcal{D}$	class	$\|\boldsymbol{x} - \boldsymbol{x}_i\|$	Rank	$\mathcal{D}'$	$\|\boldsymbol{x}' - \boldsymbol{x}'_i\|$	Rank'
(6.0,0.8)	+	2.062	4	(0.6,0.8)	0.539	5
(7.0,0.9)	+	1.166	2	(0.7,0.9)	0.608	6
(9.0,0.8)	+	1.118	1	(0.9,0.8)	0.510	4
(10.0,1.0)	+	2.119	5	(1.0,0.9)	0.728	7
(6.0,0.3)	-	2.000	3	(0.6,0.3)	0.200	1
(5.0,0.0)	-	3.015	6	(0.5,0.0)	0.424	3
(4.0,0.4)	-	4.001	7	(0.4,0.4)	0.412	2
(0,0.0.2)	-	8.001	8	(0.0,0.2)	0.808	8

Thus the classification without normalisation is that the test point comes from the positive data set, while after normalisation it comes from the negative data set.

Normalisation changes the relative weightings of components of the data. It puts each feature on the same footing. If there is no *a priori* reason to favour any feature this is likely to be a good thing to do. This would be the case if the features were measured in arbitrary units.

(5 marks)

(c) Explain what is meant by the terms *approximation error*, *estimation error* and *generalisation error*.

approximation error expresses the error associated with model mismatch.

estimation error expresses the errors associated with estimating a set of parameters from a finite amount of data.

generalisation error expresses the true error of the learning machine, which is a combination of the first two errors.

(5 marks)

(d) Describe the *kernel trick* and how it is applied in machine learning.

Usually need to build non-linear models of data to reflect the true underlying model. There are two approaches to this: a) directly build this non-linear model, b) map the data into a high dimensional space using a non-linear transformations and then solve a linear problem in that space. The kernel trick notes that often we can construct algorithms of type (b) which only involve inner products between transformed examples and as such provides a compact mathematical expression for this inner product without having to project our data into a possibly very high dimensional space. It can be employed in many algorithms which can be expressed in the form of inner products, such as KNN, SVM, ...

(5 marks)

Section B

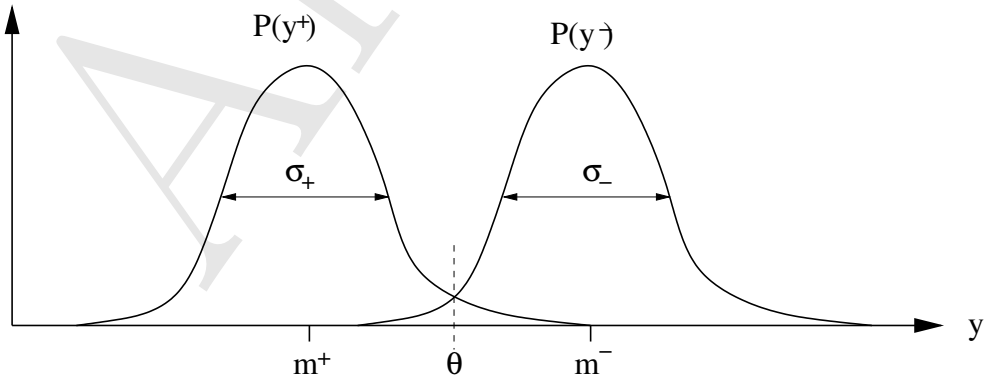
Question 2 One method (not cover in the course) for classifying a dataset into two classes is Fisher's linear discriminant (FLD) analysis. We assume that we have representative data for the two classes $\mathcal{D}^{\pm} = \{\boldsymbol{x}_i^{\pm}\}_{i=1}^{N^{\pm}}$ where the superscript '+' or '−' is used to denote the class. The goal is to find a vector $\boldsymbol{w}$ such that the quantity $y = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}$ can be used to separates the two data classes. The criteria for choosing $\boldsymbol{w}$ is to maximise the quantity

$$J(\boldsymbol{w}) = \frac{(m^+ - m^-)^2}{\sigma_+^2 + \sigma_-^2}$$

where $m^{\pm}$ denotes the means of y for the two data sets and $\sigma_{\pm}^2$ the variance then the criteria for choosing $\boldsymbol{w}$ is to maximise

(a) Assume that the distributions of $y_i^{\pm}$ in the sets $\mathcal{Y}^{\pm} = \{y_i^{\pm} = \boldsymbol{w}^{\mathsf{T}} x_i^{\pm} | x_i^{\pm} \in \mathcal{D}^{\pm}\}$ are approximately Gaussian distributed. Sketch histograms for the two sets $\mathcal{Y}^{\pm}$ showing $m^{\pm}$ and $\sigma_{\pm}$. Draw a suitable position of a threshold, θ, for making the classification decision. Explain why we normalise by the variance in the definition of $J(\boldsymbol{w})$ rather than simply maximise the difference in the two means.

This question tests their ability to apply their knowledge to a new technique.



The error in classification depends on the overlap of the tails of the two distributions. Reducing the variance reduces the size of the tails. We therefore need to play off the desire to separate the means with the desire to reduce the

width of distribution. If the distribution of the two classes depended only on the means and standard deviation then maximising $J(\boldsymbol{w})$ would maximally reduce the overlaps in the tails.

(10 marks)

(b) Show that the means and variance can be written as $m^{\pm} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}^{\pm}$ and $\sigma^2_{\pm} = \boldsymbol{w}^{\mathsf{T}}\mathbf{V}^{\pm}\boldsymbol{w}$ where $\boldsymbol{\mu}^{\pm}$ and $\mathbf{V}^{\pm}$ are the mean vectors and covariance matrices, respectively, defined by

$$\boldsymbol{\mu}^{\pm} = \frac{1}{N^{\pm}} \sum_{i=1}^{N^{\pm}} \boldsymbol{x}_i^{\pm}, \qquad \mathbf{V}^{\pm} = \frac{1}{N^{\pm} - 1} \sum_{i=1}^{N^{\pm}} (\boldsymbol{x}_i^{\pm} - \boldsymbol{\mu}^{\pm}) (\boldsymbol{x}_i^{\pm} - \boldsymbol{\mu}^{\pm})^{\mathsf{T}}$$

Tests ability to reason about data analytically. The mean is given by

$$\begin{aligned} m^{\pm} &= \frac{1}{N^{\pm}} \sum_{i=1}^{N^{\pm}} y_i^{\pm} \\ &= \frac{1}{N^{\pm}} \sum_{i=1}^{N^{\pm}} \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_i^{\pm} \\ &= \boldsymbol{w}^{\mathsf{T}} \frac{1}{N^{\pm}} \sum_{i=1}^{N^{\pm}} \boldsymbol{x}_i^{\pm} \\ &= \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\mu}^{\pm} \end{aligned}$$

(2 marks)

The variance is give by

$$\begin{aligned} \sigma^2_{\pm} &= \frac{1}{N^{\pm} - 1} \sum_{i=1}^{N^{\pm}} (y_i^{\pm} - m^{\pm})^2 \\ &= \frac{1}{N^{\pm} - 1} \sum_{i=1}^{N^{\pm}} \left(\boldsymbol{w}^{\mathsf{T}} (\boldsymbol{x}_i^{\pm} - \boldsymbol{\mu}^{\pm}) \right)^2 \\ &= \frac{1}{N^{\pm} - 1} \sum_{i=1}^{N^{\pm}} \boldsymbol{w}^{\mathsf{T}} (\boldsymbol{x}_i^{\pm} - \boldsymbol{\mu}^{\pm}) (\boldsymbol{x}_i^{\pm} - \boldsymbol{\mu}^{\pm})^{\mathsf{T}} \boldsymbol{w} \\ &= \boldsymbol{w}^{\mathsf{T}} \left(\frac{1}{N^{\pm} - 1} \sum_{i=1}^{N^{\pm}} (\boldsymbol{x}_i^{\pm} - \boldsymbol{\mu}^{\pm}) (\boldsymbol{x}_i^{\pm} - \boldsymbol{\mu}^{\pm})^{\mathsf{T}} \right) \boldsymbol{w} \\ &= \boldsymbol{w}^{\mathsf{T}} \mathbf{V}^{\pm} \boldsymbol{w} \end{aligned}$$

(5 marks)

TURN OVER

(c) Show that the vector $\boldsymbol{w}^*$ which maximises $J(\boldsymbol{w})$ satisfies $\boldsymbol{w}^* \propto (\mathbf{V}^+ + \mathbf{V}^-)^{-1}(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)$

Tests ability to carry out a typical calculation for finding an optimal training vector. The optimal vector w^* will satisfy the equation $\boldsymbol{\nabla} J(\boldsymbol{w}) = 0$. Now

$$\begin{aligned}
\boldsymbol{\nabla} J(\boldsymbol{w}) &= \boldsymbol{\nabla} \frac{(m^+ - m^-)^2}{\sigma_+^2 + \sigma_-^2} \\
&= \boldsymbol{\nabla} \frac{\left(\boldsymbol{w}^\mathsf{T}(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)\right)^2}{\boldsymbol{w}^\mathsf{T}(\mathbf{V}^+ + \mathbf{V}^-)\boldsymbol{w}} \\
&= 2\frac{\left(\boldsymbol{w}^\mathsf{T}(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)\right)(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)}{\boldsymbol{w}^\mathsf{T}(\mathbf{V}^+ + \mathbf{V}^-)\boldsymbol{w}} - 2\frac{\left(\boldsymbol{w}^\mathsf{T}(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)\right)^2}{\left(\boldsymbol{w}^\mathsf{T}(\mathbf{V}^+ + \mathbf{V}^-)\boldsymbol{w}\right)^2}(\mathbf{V}^+ + \mathbf{V}^-)\boldsymbol{w}
\end{aligned}$$

setting $\boldsymbol{\nabla} J(\boldsymbol{w}) = 0$ we find

$$\boldsymbol{w} = c(\mathbf{V}^+ + \mathbf{V}^-)^{-1}(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)$$

where

$$c = \frac{\boldsymbol{w}^\mathsf{T}(\mathbf{V}^+ + \mathbf{V}^-)\boldsymbol{w}}{\boldsymbol{w}^\mathsf{T}(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)}$$

Because c is proportional to $\|w\|$ the norm of w is not determined.

(7 marks)

(d) Discuss the similarities and differences between Fisher's Linear Discriminant and a perceptron.

Test understanding of the relation between different learning machines. FLD has the same structure as a perceptron. The difference is in how the weight vector w is chosen. In FLD the weights are chosen assuming that the data is Gaussianly distributed. In the perceptron individual data points are taken into account. In particular the data in the tails of the distribution determine the position of the separating plane. In contrast, the bulk properties (first and second moments) determine the position of the separating plane in FLD.

(3 marks)

Question 3

(a) Describe the similarities and difference between multi-layer perceptrons (MLPs), radial basis function networks (RBFs) and support vector machines (SVMs).

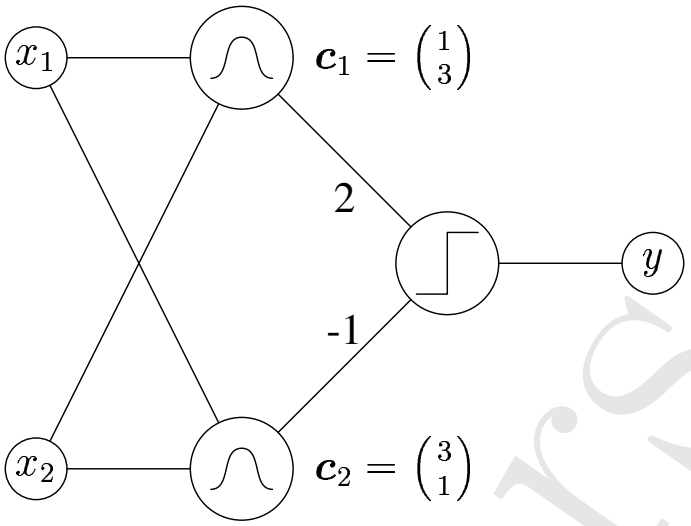
Test non-mathematical knowledge. Many ways to answer. The points I am looking for are

- **All three techniques based of the perceptron. In MLPs the earlier layers are perceptrons, in RBFs they are radial basis functions and in SVMs they are the features corresponding to the eigenvalues of a kernel.**
- **All three can be used for regression of classification.**
- **MLPs trained using back-propagation of errors. They have a non-unique solution. Complexity depends on number of hidden nodes. Liable to over-fit the training data. Often use *ad hoc* methods such as early stopping to prevent over-fitting. Can have many output nodes.**
- **RBFs typically use unsupervised learning to choose the centres for the input layer. The labelled data used to train the final layer (a perceptron). Training is fast. Can have many output nodes. Often use regulariser on the output layer. The solution found is unique.**
- **SVMs use a kernel function to perform a mapping into a very high dimensional feature space. An optimally stable perceptron is used in the feature space. This controls the capacity of the learning machine reducing the problem of over-fitting. The learning algorithm uses quadratic optimisation. The computation complexity grows as the number of training patterns cubed. For very large datasets SVMs can become impractical. The solution found is unique.**

(15 marks)

(b) The diagram below shows a radial basis function with two inputs and two centres at $\boldsymbol{c}_1 = (1, 3)^\mathsf{T}$ and $\boldsymbol{c}_2 = (3, 1)^\mathsf{T}$ respectively. The radial basis function are Gaussians of the form $\exp(-\|\boldsymbol{x} - \boldsymbol{c}_i\|^2/2)$. The final layer is a 0-1 step perceptron with zero bias. The weights from the radial basis function to the perceptron are $w_1 = 2$ and $w_2 = 1$. Write down a formula describing the output of the network as a function of the inputs and sketch the approximate position of the separating surface in the plane $\boldsymbol{x} = (x_1, x_2)$.

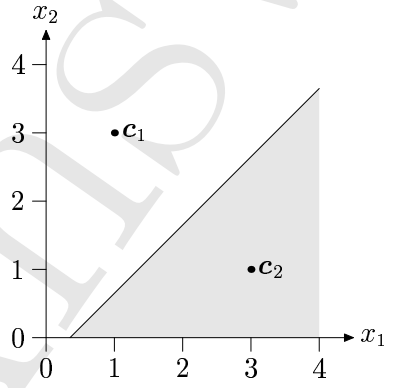
TURN OVER



Test ability to work out explicit details for a simple RBF network. The output of the network is

$$y = \Theta\left(2\exp(-\|\boldsymbol{x} - (1,3)^{\mathsf{T}}\|^2/2) - \exp(-\|\boldsymbol{x} - (3,1)^{\mathsf{T}}\|^2/2)\right)$$

where $\Theta(x)$ represents the step function. The separating surface is actually given by $x_2 = x_1 - \log(2)/2$, although any surface which is vaguely in the right place is sufficient to get full marks.



(10 marks)

Section C

Question 4

(a) Explain what is meant by the term over-parameterisation. State how it is removed from the hyperplane, $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b = 0$, in the linear Support Vector Machine formulation to produce a canonical hyperplane.

Over-parameterisation refers to case when the parameters in an equation are not independent, and hence two different sets of parameters can describe an identical solution. The over-parameterisation in $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b = 0$ is removed by adding the constraint $\min_i \left|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b\right| = 1$ where $\boldsymbol{x}_i$ are the input space coordinates of the training examples.

(3 marks)

(b) State the condition for separability of the two-class data-set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n}, \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ with this canonical hyperplane.

$y_i \left[\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b\right] \geq 1, \quad i = 1, \ldots, n.$

(3 marks)

(c) State the maximum margin principle and derive an expression for the Lagrangian of the resulting optimisation problem.

The maximum margin principle states: "The set of vectors, $\boldsymbol{x}_i$, is said to be optimally separated by the hyperplane if it is separated without error and the distance between the closest vector to the hyperplane is maximal."

The distance $d(\boldsymbol{w}, b; \boldsymbol{x})$ of a point $\boldsymbol{x}$ from the hyperplane $(\boldsymbol{w}, b)$ is,

$$d(\boldsymbol{w}, b; \boldsymbol{x}) = \frac{\left|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b\right|}{\|\boldsymbol{w}\|}.$$

The optimal hyperplane is given by maximising the margin, ρ, subject to the separability constraints of b). The margin is given by,

$$\begin{aligned}
\rho(\boldsymbol{w}, b) &= \min_{\boldsymbol{x}_i : y_i = -1} d(\boldsymbol{w}, b; \boldsymbol{x}_i) + \min_{x_i : y_i = 1} d(\boldsymbol{w}, b; \boldsymbol{x}_i) \\
&= \min_{\boldsymbol{x}_i : y_i = -1} \frac{\left|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b\right|}{\|\boldsymbol{w}\|} + \min_{\boldsymbol{x}_i : y_i = 1} \frac{\left|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b\right|}{\|\boldsymbol{w}\|} \\
&= \frac{1}{\|\boldsymbol{w}\|} \left(\min_{\boldsymbol{x}_i : y_i = -1} \left|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b\right| + \min_{\boldsymbol{x}_i : y_i = 1} \left|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b\right| \right) \\
&= \frac{2}{\|\boldsymbol{w}\|}
\end{aligned}$$

Hence the hyperplane that optimally separates the data is the one that minimises

$$\Phi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2.$$

subject to $y^i \left[\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right] \geq 1$ **and hence the Lagrangian is,**

$$\Phi(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \tfrac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left[\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b\right] - 1\right), \quad \alpha_i \geq 0.$$

(10 marks)

(d) Solve the Lagrangian problem, $\max_{\boldsymbol{\alpha}} \left(\min_{\boldsymbol{w},b} \Phi(\boldsymbol{w}, b, \boldsymbol{\alpha})\right)$, to show that the solution for the Lagrange multipliers can be written as a quadratic program.

The minimum with respect to $\boldsymbol{w}$ **and** b **of the Lagrangian,** Φ**, is given by,**

$$\frac{\partial \Phi}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial \Phi}{\partial \boldsymbol{w}} = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i.$$

Hence, by substitution and rearrangement the dual problem is,

$$\max_{\boldsymbol{\alpha}} -\tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\mathsf{T} \boldsymbol{x}_j + \sum_{k=1}^{n} \alpha_k,$$

and hence the solution to the problem is given by,

$$\min_{\boldsymbol{\alpha}} \tfrac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\mathsf{T} \boldsymbol{x}_j - \sum_{k=1}^{n} \alpha_k,$$

with constraints,

$$\alpha_i \geq 0 \quad i = 1, \ldots, n$$

$$\sum_{j=1}^{n} \alpha_j y_j = 0.$$

This is equivalent to the quadratic program formulation by noting $H_{i,j} = y_i y_j \boldsymbol{x}_i^\mathsf{T} \boldsymbol{x}_j$ **and** $c_i = -1$**.**

(9 marks)

Question 5

(a) Describe the difference between *experimental* and *observational* data, with reference to examples.

---

experimental **gives us control over the way the data is collected and can hence collect points from any part of the input space, e.g. scientific experiment.** ***observational*** **data gives us no control over where the data points will lie, e.g. Google analysis of trawled web pages.**

(6 marks)

(b) Describe the *curse of dimensionality* and its consequences.

Data required grows exponentially with features. Only way to limit this is to introduce priors.

(5 marks)

(c) Describe the difference between *feature selection* and *feature extraction.*

---

feature selection means to remove redundant features which convey little or no information to the learning task. Feature extraction is concerned with locating new features which describe an "interesting" subspace of the data, e.g. PCA.

(6 marks)

(d) As a machine learning expert you are asked to consult for a company who are considering apply machine learning techniques to their data. Describe how you would go about this.

Problem: They don't understand ML, you don't understand their problem/data. Here is one approach: You need some sort of cyclic process to boost each others understanding. Remember they probably have many misconceptions about ML. Setup a meeting where you select two or three representative learning algorithms to get across the principles of learning (not the details). They should come out knowing the difference between density estimation, classification, regression supervised/unsupervised learning and off/online learning, as well as the scale/time complexities of common algorithms. Get them to introduce their problem. What attributes do the features have? numeric, text,... What sort of outputs are required? probalilistic? How much data is available? How expensive is it to collect? What is the most appropriate way to measure the error? Are costs balanced? Identify a small dataset that exists to explore some of the potential solutions and present at the next meeting...

(8 marks)

END OF PAPER