Informatics Board of Studies - Course Proposal

Proposed course title: Introduction to Data Analytics and Machine Learning

Proposer(s): Michelle Galea, Maria Wolters, Gary McGilvray, Dave Robertson

Date: 8/01/15

This template contains the following sections, which should be prepared roughly in the order in which they appear (to avoid spending too much time on preparation of proposals that are unlikely to be approved):

1. **Case for Support**
   
   - to be supplied by the proposer and shown to the BoS Academic Secretary prior to preparation of an in-depth course description

   1a. **Overall contribution to teaching portfolio**
   
   1b. **Target audience and expected demand**
   
   1c. **Relation to existing curriculum**
   
   1d. **Resources**

2. **Course descriptor**
   
   - this is the official course documentation that will be published if the course is approved, ITO and the BoS Academic Secretary can assist in its preparation

3. **Course materials**
   
   - these should be prepared once the Board meeting at which the proposal will be discussed has been specified

   3a. **Sample exam question**
   
   3b. **Sample coursework specification**
   
   3c. **Sample tutorial/lab sheet question**
   
   3d. **Any other relevant materials**

4. **Course management**
   
   - this information can be compiled in parallel to the elicitation of comments for section 5.

   4a. **Course information and publicity**
   
   4b. **Feedback**
   
   4c. **Management of teaching delivery**

5. **Comments**
   
   - to be collected by the proposer in good time before the actual BoS meeting and included as received

   5a. **Year Organiser Comments**
   
   5b. **Degree Programme Co-Ordinators**
   
   5c. **BoS Academic Secretary**


1. **Case for Support**

   [This section should summarise why the new course is needed, how it fits with the existing course portfolio, the curricula of our Degree Programmes, and delivery of teaching for the different years it would affect.]

   The proposed course in Data Analytics/Applied Machine Learning is to be offered as an online distance learning course, in order to take advantage of the active and growing market in distance education in Data Science.

   This course is heavily based on an existing one offered by the school – Introductory Applied Machine Learning (IAML), INFR02029. The distance learning offering will allow interested students to study in-
demand topics without the constraints of campus attendance, and School staff to develop skills and experience in the development and delivery of distance education programmes, and in the support of distance education students. It will also provide the School with another opportunity to establish new relationships across industry, the public sector and academia.

1a. Overall contribution to teaching portfolio
[Explain what motivates the course proposal, e.g. an emergent or maturing research area, a previous course having become outdated or inappropriate in other ways, novel research activity or newly acquired expertise in the School, offerings of our competitors.]

The School is currently in a position to take advantage of the growing market in distance education, specifically in the area of Data Science, and compete with other educational institutions in the UK and abroad offering similar services. The target audience is professionals working in a data-intensive environment, and the School can capitalise on existing links with a wide range of small and large businesses and government agencies with which it collaborates through its various initiatives, the commercialisation unit, the Data Science centre for doctoral training, and other relevant innovation centres.

1b. Target audience and expected demand
[Describe the type of student the course would appeal to in terms of background, level of ability, and interests, and the expected class size for the course based on anticipated demand. A good justification would include some evidence, e.g. by referring to projects in an area, class sizes in similar courses, employer demand for the skills taught in the course, etc.]

Students are expected to be primarily professionals in data-intensive industries looking to extend their expertise.

Market research in early 2014 (conducted as a precursor to obtaining the University’s support for our programme through the Distance Education Initiative) examined industry activities relating to Data Science. Below is a summary of the survey results relevant to distance education in the area:

- 69 responses from a number of public and private sector organisations.
- 45% of respondents would look to engage with a university to further knowledge/skills.
- Companies spanned a number of industries: Aerospace, Defence, Chemicals, Creative Industry, Education, Engineering, Energy, Financial Services and others.
- Only 1% of respondents had previously worked with a Scottish university through CPD.
- But 63% of respondents either said they were Somewhat Interested, Interested or Very Interested in such Continued Professional Development/Training in the future.

2011/12 HESA data indicates that the UK Informatics postgraduate sector is large, with over 16,000 students, and that there is an active area for online distance learning with 183 students.

1c. Relation to existing curriculum
[This section should describe how the proposed course relates to existing courses, programmes, years of study, and specialisms. Every new course should make an important contribution to the delivery of our Degree Programmes, which are described at http://www.drps.ed.ac.uk/12-13/dpt/drps_inf.html. Please name the Programmes the course will contribute to, and justify its contribution in relation to courses already available within those programmes. For courses available to MSc students, describe which specialism(s) the course should be listed under (see http://www.inf.ed.ac.uk/student-services/teaching-organisation/taught-course-information/year-guides/taught-postgraduate-year-guide/degree-requirements/specialist-areas), and what its significance for the specialism would be. Comment on the fit of the proposed course with the structure of academic years for which it should be offered. This is}
The course is required as a core module to the PG Certificate in Data Science (by Distance Education), with a planned launch date of August 2015, and to the MSc in Data Science (by Distance Education), with a planned launch date of 2016. We also intend to allow individual courses to be taken for those interested in comparatively modest investment in continuous professional development.

1d. Resources

[While course approvals do not anticipate the School's decision that a course will actually be taught in any given year, it is important to describe what resources would be required if it were run. Please describe how much lecturing, tutoring, exam preparation and marking effort will be required in steady state, and any additional resources that will be required to set the course up for the first time. Please make sure that you provide estimates relative to class size if there are natural limits to its scalability (e.g. due to equipment or space requirements). Describe the profile of the course team, including lecturer, tutors, markers, and their required background. Where possible, identify a set of specific lecturers who have confirmed that they would either like to teach this course apart from the proposer, or who could teach the course in principle. It is useful to include ideas and suggestions for potential teaching duty reallocation (e.g. through course sharing, discontinuation of an existing course, voluntary teaching over and above normal teaching duties) to be taken into account when resourcing decisions are made.]

Course lectures: Slides and videos for lectures of the existing IAML course may be reused, minimising course development and preparation. However, moving these onto the online distance learning delivery platform with additional, with the creation of additional resources as necessary will require some effort. We expect 3-4 person weeks are required to convert existing material and create new material as necessary, but funding for that already exists and will not impact on Informatics staff.

Online tutorials: We expect these to be run by current or new tutors, mainly current PhD students, especially in the first few years when student numbers will be at their lowest.

Coursework marking effort: we aim for assignments and exams to be marked by members of the course team, or via PhD students, especially in the first few years when student numbers may be at their lowest.

Student interaction: it is important that distance learning students feel a part of both the School and University. Resources are therefore required to engage with and encourage them to work closely with the course team, and to link to peer-to-peer support via the Moodle and School social platforms, discussion forums, and specific course/session levels.

Coursework resources: Aridhia have agreed work as part of the course production team to produce practical exercises that can be run on-line via their analytics toolkit. University IS have agreed to support the toolkit on its servers and are dealing with contractual arrangements. It is also likely that we may wish to use WEKA (currently running on DICE machines for the IAML course), particularly if we want to reuse practical introductory examples from IAML. If so, we will negotiate with IS to support WEKA, so as to avoid any additional support burden on Informatics.

Exam preparation: We intend, at least in the first run of the course, to follow closely the style and content of IAML exams. These will be prepared in collaboration with current IAML organizers.

Course team: As part of the wider Data Science effort, a core team currently exists to support various aspects related to all distance education courses, e.g. manage underlying distance education platforms, provide administration, etc. One course tutor would be required to answer student questions, mark or manage marking of course assignments, provide feedback, etc.

2. Course descriptor
Course Title: Introduction to Data Analytics and Machine Learning.

SCQF Credit Points: 10
SCQF Credit Level: 11

The current IAML course is a level 9, however, we plan to add assessed material to some of the practical exercises that will challenge the depth of a student’s understanding. This additional challenge will be enabled and supplemented by online discussions on specific topics from the syllabus.

[These levels correspond to different levels of skills and outcomes, see http://www.scqf.org.uk/The%20Framework/Level%20Descriptors. At University level, Year 1/2 courses are normally level 8, Year 3 can be level 9 or 10, Year 4 10 or 11, and Year 5/MSc have to be level 11. MSc programmes may permit a small number (up to 30 credits overall) of level 9 or 10 courses.]

Normal Year Taken: MSc
Also available in years: n/a

Subject Area and Specialism Classification:
Learning from Data; Knowledge Management, Representation and Reasoning

Appropriate/Important for the Following Degree Programmes:
This course is a core module for the online distance education PG Certificate in Data Science (launch August 2015), and the online distance education MSc in Data Science (launch 2016).

Timetabling Information:
[Provide details on the semester the course should be offered in, specifying any timetabling constraints to be considered (e.g. overlap of popular combinations, other specialism courses, external courses etc).]
To be determined as part of our distance education marketing.

School Acronym: INF-???
[This can be provided by the Informatics Teaching Organisation.]

Short Course Description:
Organisations seek to make better decisions by examining their data with an aim to discovering and/or drawing conclusions about the information contained within. This course is about the principled application of machine learning techniques to extracting information from data. The main area that will be discussed is supervised learning, which is concerned with learning to predict an output, given inputs. A second area of study is unsupervised learning, where we wish to discover the structure in a set of patterns, i.e. there is no output "teacher signal". The primary aim is to provide the student with a set of practical tools that can be applied to solve real-world problems in machine learning, coupled with an appropriate, principled approach to formulating a solution.

Pre-Requisite Courses:
EPCC is preparing a practical, basic introductory course on data science (Practical Data Science) which has been approved through their Board of Studies and would be a good precursor to our analytics course but we do not want to make Practical Data Science a strict pre-requisite.
Co-Requisite Courses: none

Prohibited Combinations: none

Other Requirements:
Maths requirements:
4. Special functions: Log, exp
5. Geometry: Basics of lines, planes and hyperplanes. Coordinate geometry of circle, sphere, ellipse, ellipsoid and n-dimensional generalizations.
6. Entropy: is useful, but will be covered in the lectures.

Programming requirements: None beyond some basic scripting for the Aridhia toolset (we will balance this against WEKA examples if we wish to reduce the “programming” component to near zero).

Available to Visiting Students: Yes/No
[Provide a justification if the answer is No.]
Potentially, this course could be made available to all appropriate students at UoE, although it is being created for a distance learning market.

Summary of Intended Learning Outcomes:
A student who has successfully completed this course should be able to:
1. Explain the scope, goals and limits of machine learning, and the main sub-areas of the field.
2. Describe the various techniques covered in the syllabus and where they fit within the structure of the discipline.
3. Students should be able to critically compare, contrast and evaluate the different machine learning techniques in terms of their applicability to different data exploration problems.
4. Given a data set and problem students should be able to use appropriate software to apply these techniques to the data set to solve the problem.
5. Given appropriate data students should be able to use a systematic approach to conducting experimental investigations and assessing scientific hypotheses.

Assessment Information
[Provide a description of all types of assessment that will be used in the course (e.g. written exam, oral presentation, essay, programming practical, etc) and how each of them will assess the intended learning outcomes listed above. Where coursework involves group work, it is important to remember that every student has to be assessed individually for their contribution to any jointly produced piece of work. Provide information on how long students can expect to spend on assignments.]

Assessment may include use of online questionnaires and progress quizzes, peer assessment, written assignments and final course exam. Each learning outcome will be tested by at least one of the assessment methods.

Assessment Weightings:
<table>
<thead>
<tr>
<th>Assessment</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written Examination</td>
<td>70%</td>
</tr>
<tr>
<td>Practical Examination</td>
<td>0%</td>
</tr>
<tr>
<td>Coursework</td>
<td>30%</td>
</tr>
</tbody>
</table>
Time spent on assignments: 25 hours

[Weightings up to a 70/30 split between exam and coursework are considered standard, any higher coursework percentage requires a specific justification. The general expectation is that a 10-point course will have an 80/20 split and include the equivalent of one 20-hour coursework assignment (although this can be split into several smaller pieces of coursework. The Practical Examination category should be used for courses with programming exams. You should not expect that during term time a student will have more than 2-4 hours to spend on a single assignment for a course per week. Please note that it is possible, and in many cases desirable, to include formative assignments which are not formally assessed but submitted for feedback, often in combination with peer assessment.]

Academic description:
[A more technical summary of the course aims and contents. May include terminology and technical content that might be more relevant to colleagues and administrators than to students.]

Syllabus:

Relevant QAA Computing Curriculum Sections:
[Please see http://www.qaa.ac.uk/Publications/InformationAndGuidance/Pages/Subject-benchmark-statement-Computing.aspx to check which section the course fits into.]

Artificial Intelligence, Human-Computer Interaction (HCI), Intelligent Information Systems Technologies, Natural Language Computing, Simulation and Modelling, Theoretical Computing

Transferrable skills: n/a

Reading List:

Study Abroad: n/a

Breakdown of Learning and Teaching Activities:

20 lecture hours and 8 tutorial/studio hours each week, with 4 coursework assignments.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture Hours</td>
<td>20</td>
</tr>
<tr>
<td>Seminar/Tutorial Hours</td>
<td>4</td>
</tr>
<tr>
<td>Supervise practical/Workshop/Studio</td>
<td>4</td>
</tr>
<tr>
<td>Summative assessment hours</td>
<td>2</td>
</tr>
<tr>
<td>Feedback/Feedforward hours</td>
<td>2</td>
</tr>
<tr>
<td>Directed Learning and Independent</td>
<td>68</td>
</tr>
<tr>
<td>Learning hours</td>
<td>100</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>
The breakdown of learning and teaching activities should only include contact hours with the students; everything else should be accounted for in the Directed Learning and Independent Learning hours. The total being $10 \times$ course credits. Assume 10 weeks of lectures slots and 10 weeks of tutorials, though not all of these need to be filled with actual contact hours. As a guideline, if a 10-pt course has 20 lecture slots in principle, around 15 of these should be filled with examinable material, the rest should be used for guest lectures, revision sessions, introductions to assignments, etc. Additional categories of learning and teaching activities are available, a full list can be found at:


Keywords:
Data analytics, machine learning, data mining

3. Course materials
3a. Sample exam question(s)
[Sample exam questions with model answers to the individual questions should be provided. A justification of the exam format should be provided where the suggested format non-standard. The online list of past exam papers gives an idea of what exam formats are most commonly used and which alternative formats have been http://www.inf.ed.ac.uk/teaching/exam_papers/.

Exam papers will be similar to IAML exams.

3b. Sample coursework specification
[Provide a description of a possible assignment with an estimate of effort against each sub-task and a description of marking criteria.]

Assignments will be similar to existing IAML courses assignments, though some will present additional material to challenge a student’s more in-depth understanding of the topic.
http://www.inf.ed.ac.uk/teaching/courses/iaml/2014/assts/asst1/a112_1.pdf

3c. Sample tutorial/lab sheet questions
[Provide a list of tutorial questions and answers and/or samples of lab sheets.]

Tutorial and lab sheets will be similar to existing IAML material:
http://www.inf.ed.ac.uk/teaching/courses/iaml/tut/tutorial-1.pdf
http://www.inf.ed.ac.uk/teaching/courses/iaml/lab/lab1.html

3d. Any other relevant materials
[Include anything else that is relevant, possibly in the form of links. If you do not want to specify a set of concrete readings for the official course descriptor, please list examples here.]

Recommended reading:
• Other good books:
  ◦ Pattern Recognition and Machine Learning by C. Bishop
  ◦ Elements of Statistical Learning by Hastie, Tibshirani and Friedman
  ◦ Bayesian Reasoning and Machine Learning by D. Barber
  ◦ Machine Learning by T. Mitchell
  ◦ Reinforcement Learning by R. Sutton and A. Barto
• Past years' exam papers are available from 2009 and 2010.
• The lecture notes from the old Learning from Data course are useful, although they contain more mathematical detail than we are expecting for IAML.

A Few Useful Things to Know about Machine Learning by P. Domingos

4. Course management
4a. Course information and publicity
[Describe what information will be provided at the start of the academic year in which format, how and where the course will be advertised, what materials will be made available online and when they will be finalised. Please note that University and School policies require that all course information is available at the start of the academic year including all teaching materials and lecture slides.]

Course content will be served from Moodle hosted by the University, initially hosting the lecture set and reading list. The current Introductory Applied Machine Learning (IAML) lecture set can be used, perhaps subject to minor modifications. Similarly, assignments may be taken from the IAML course, and these can be added as the student progresses. This course will be advertised alongside our marketing of the PG Certificate in Data Science programme, which will include Informatics website advertising, making relevant industries aware, etc. We have dedicated administrative support for this in the first year, funded by the Distance Education Initiative.

4b. Feedback
[Provide details on feedback arrangements for the course. This includes when and how course feedback is solicited from the class and responded to, what feedback will be provided on assessment (coursework and exams) within what timeframe, and what opportunities students will be given to respond to feedback. The University is committed to a baseline of principles regarding feedback that we have to implement at every level, these are described at http://www.docs.sasg.ed.ac.uk/AcademicServices/Policies/Feedback_Standards_Guiding_Principles.pdf. Further guidance is available from http://www.enhancingfeedback.ed.ac.uk/staff.html.]

We plan a rich and well-resourced level of engagement between distance education students and world-leading Informatics teaching and research staff:

• Extensive use of the School and University level virtual learning environments (VLE) and social platform is planned.
• Course forums will allow students to ask questions to both teaching staff and to other students.
• Virtual class presentation and project spaces will be available based on the extensive experience of the School with the Virtual University of Edinburgh (Vue) facilities.

The course tutor will comment on each student’s work either via Moodle or a VLE.

4c. Management of teaching delivery
[Provide details on responsibilities of each course staff member, how the lecturer will recruit, train, and supervise other course staff, what forms of communication with the class will be used, how required equipment will be procured and maintained. Include information about what support will be required for this from other parties, e.g. colleagues or the Informatics Teaching Organisation.]

We expect that the course tutor needs to provide support to course students and flag any issues that do arise related to the delivery of the course. As this course is part of the Data Science Distance Education
effort, the Data Science team or University support teams will handle most issues concerned with content delivery, e.g. issues with university-hosted software or Moodle. Communication with the class will primarily occur via Moodle and a course mailing list however at times, the University level VLE’s will also be used; all of which are supported by the relevant teams. Minimal support is required from the ITO.

5. Comments
[This section summarises comments received from relevant individuals prior to proposing the course.]

This is one of the first of a new line of distance education courses to be run within a cross-College framework so the concepts of year organiser and degree programme coordinator (though still applicable) are not the same as for our residential courses. We are in discussions with our Assistant Principal Learning and Development to build an appropriate “container” for these courses with associated academic and administrative responsibilities.

5a. Year Organiser Comments
[Year Organisers are responsible for maintaining the official Year Guides for every year of study, which, among other things, provide guidance on available course choices and specialist areas. The Year Organisers of all years for which the course will be offered should be consulted on the appropriateness and relevance on the course. Issues to consider here include balance of course offerings across semesters, subject areas, and credit levels, timetabling implications, fit into the administrative structures used in delivering that year.]

5b. Degree Programme Co-Ordinators
[Degree Programme Co-Ordinators are responsible for maintaining the official Degree Programme Specifications and Degree Programme Table for a given subject area which, among other things, specify the content of courses taken in a Degree Programme. The Degree Programme Co-Ordinators of the relevant subject areas that the course is proposed for should comment on the fit with the current curriculum of the relevant Degree Programmes. Issues to consider here are dependencies arising from pre-, co-requisites, and forbidden combinations, balance of different topics in a Degree Programme, etc.]

5c. BoS Academic Secretary
[Any proposal has to be checked by the Secretary of the Board of Studies prior to discussion at the actual Board meeting. This is a placeholder for their comments, mainly on the formal quality of the content provided above.]