Key Course Selection for Academic Early Warning Based on Gaussian Processes

Min Yin [*], Jing Zhao [⋆], and Shiliang Sun

Department of Computer Science and Technology, East China Normal University
500 Dongchuan Road, Shanghai 200241, P. R. China
jzhao2011@gmail.com slsun@cs.ecnu.edu.cn

Abstract. Academic early warning (AEW) is very popular in many colleges and universities, which is to warn students who have very poor grades. The warning strategies are often made according to some simple statistical methods. The existing AEW system can only warn students, and it does not make any other analysis for academic data, such as the importance of courses. It is significant to discover useful information implicit in data by some machine learning methods, since the hidden information is probably ignored by the simple statistical methods. In this paper, we use the Gaussian process regression (GPR) model to select key courses which should be paid more attention to. Specifically, an automatic relevance determination (ARD) kernel is employed in the GPR model. The length-scales in the ARD kernel as hyperparameters can be learned through the model selection procedure. The importance of different courses can be measured by these corresponding length-scales. We conduct experiments on real-world data. The experimental results show that our approaches can make reasonable analysis for academic data.

Keywords: Academic early warning, key course selection, Gaussian process regression, automatic relevance determination kernel

1 Introduction

Many colleges and universities are working on academic early warning (AEW) which can give prompt warnings to the students who have poor test scores. AEW is among the recent computational education problems discussed in Sun [8]. Effective warning strategies will promote the students' learning or improve their learning methods. In most colleges and universities, the existing warning strategies are made according to some simple statistical methods. For example, the AEW system will send warning letters to the students who have more than ten failed credits in a semester. However, this simple statistical method can only discover the explicit information appears in the data. Using machine learning methods to model the data can excavate some hidden but useful information implicit in the data. Further, the existing AEW system can only make some warning specific to students [3]. It does not make any analysis about the courses.

[⋆] The authors contributed equally to this work.

Some useful advice on the courses will help the students to study purposefully and effectively. Therefore, selecting key courses by machine learning methods is significant and has a practical value.

Key course selection can be regarded as feature selection in the machine learning area. There are several popular methods for feature selection. For example, Lasso regression [10, 1, 12] and $\mathcal{L}_1$-norm support vector regression [13, 5] both use $\mathcal{L}_1$-norm regularization to achieve sparse feature selection. The selection results from the above two methods are deterministic. Some probabilistic models based on the Bayesian framework can provide uncertainty estimates for features. The importance of the features can be presented by the weight ratios of each features. The features with much higher weight ratios are more likely to be the key features. In allusion to the key course selection in the AEW system, given the weight ratio of each course, users can filter the key courses depending on the practical demands.

When it comes to probabilistic models, the Gaussian process regression (GPR) model is a typical probabilistic model [4]. The GPR model provides a flexible framework for probabilistic regression and classification. It is widely used to solve the nonlinear regression problems attributed to its elegant formulation [4, 2]. Several improved approaches for the GPR model are successively put forward, such as sparse Gaussian process [11, 14] and mixtures of Gaussian processes [9, 7]. Besides the elegant nonlinear modeling form, the flexibility of choosing kernel functions is another attractive feature of GPR models. Some special data characteristics can be captured by the particular kernel functions. For example, the automatic relevance determination (ARD) kernel is able to capture the importance of different features. The ARD kernel has been used successfully for removing irrelevant features [4]. The hyperparameters introduced by kernel functions can be adaptively learned by model selection methods.

In this paper, we use a GPR model with an ARD kernel to select key courses which should be paid more attention to. The length-scales as hyperparameters in the ARD kernel can be learned through the model selection procedure. The importance of different courses can be measured by these length-scales. We conduct experiments on real-world data. Due to the practical situation that different students sometimes choose different courses, the collected data need to be reconstructed. After reconstructing the data by the nearest neighbor data-filling algorithm, we use the GPR model with an ARD kernel to model the reconstructed data and select the key courses.

2 Gaussian Process Regression Model

In this section, we will introduce the GPR model from the function-space view, and analyze the model selection methods for the GPR model.

2.1 Gaussian Process Regression Model

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution. From the perspective of function space, Gaussian process can be seen as a distribution of function. The characteristic of Gaussian process is determined by mean function and covariance function. Define mean function $m(\mathbf{x})$ and the covariance function $\kappa(\mathbf{x}, \mathbf{x}')$ of a Gaussian process $f(\mathbf{x})$ as

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \tag{1}$$

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \tag{2}$$

The Gaussian process $f(\mathbf{x})$ can be written as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')). \tag{3}$$

In most cases, we can only get access to noisy versions thereof $y = f(\mathbf{x}) + \epsilon$. Assuming additive independent identically distributed Gaussian noise ϵ with variance σ_n^2, the joint distribution of the observed values $\mathbf{y}$ and the test outputs $\mathbf{f}_*$ is

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right), \tag{4}$$

where K is the covariance matrix calculated by the kernel function $\kappa(\mathbf{x}, \mathbf{x}')$.

2.2 Model Selection

The GPR model can use different covariance functions. Squared exponential covariance function is a common covariance function. Furthermore, by using the ARD squared exponential kernel, the model selection procedure allows us to automatically infer the importance of the input features without introducing explicit regularization. For the purpose of doing automatic model selection of the dimensionality of latent space, the kernel can be chosen to follow the ARD squared exponential form:

$$\kappa(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\{-\frac{1}{2}\sum_{d=1}^{D} \frac{1}{\ell_d^2}(x_d - x_d{}')^2\}, \tag{5}$$

where ℓ_d is the length-scale of the covariance and σ_f^2 is the signal variance. σ_n^2 mentioned in (4) is the noise variance. In general we call these free parameters hyperparameters. We use symbol $\boldsymbol{\theta}$ to denote the hyperparameters in the Gaussian process regression model, i.e., $\boldsymbol{\theta} = \{\{\ell_d^2\}, \sigma_f^2, \sigma_n^2\}$. Such a covariance function implements automatic relevance determination, since the inverse of the length-scale determines how relevant an input feature is. We will introduce two kinds of model selection methods for the GPR model. One is the Type II maximum likelihood and the other one is maximizing a posteriori [4, 6].

Type II Maximum Likelihood In Type II Maximum Likelihood (ML-II), one needs to calculate the negative logarithmic marginal likelihood of the samples, $L(\boldsymbol{\theta}) = -\log p(\mathbf{y}|X, \boldsymbol{\theta})$, and then calculate the partial derivatives of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ [4]. Through the model selection procedure, the hyperparameters in the ARD kernel which represent the importance of the features can be automatically determined. In our practical problems, the key courses can be chosen according to the ratios of the magnitudes of the inverse length-scales. We denote these ratios as the weight ratios of courses.

Maximizing a Posteriori In maximizing a posteriori (MAP), one needs to compute the posterior of the hyperparameters which is expressed as

$$p(\boldsymbol{\theta}|\mathbf{y}, X, \mathcal{H}_p) = \frac{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_p) p(\boldsymbol{\theta}|\mathcal{H}_p)}{p(\mathbf{y}|X, \mathcal{H}_p)}, \tag{6}$$

where $p(\boldsymbol{\theta}|\mathcal{H}_p)$ is the prior for the hyperparameters named as hyper prior. $\mathcal{H}_p$ represents the parameters in the hyper prior distribution, which can be set by hand according to actual situations.

In our experimental settings, we use Gamma hyper prior for the inverse length-scales, and use Gaussian hyper prior for both the logarithmic signal variance and logarithmic noise variance. The Gamma hyper prior for the inverse length-scale $\frac{1}{\ell_d}$ is expressed as

$$\frac{1}{\ell_d} \sim \text{Gamma}(\alpha, \lambda), \tag{7}$$

where the expectation and variance of the Gamma hyper prior are

$$\mathbb{E}(\frac{1}{\ell_d}) = \frac{\alpha}{\lambda}, \quad \mathbb{V}(\frac{1}{\ell_d}) = \frac{\alpha}{\lambda^2}. \tag{8}$$

The Gaussian hyper prior of the logarithmic signal and noise variance are expressed as

$$\log(\sigma_f^2) \sim \mathcal{N}(\mu_0, \sigma_0), \quad \log(\sigma_n^2) \sim \mathcal{N}(\mu_1, \sigma_1). \tag{9}$$

Given the above Gamma hyper prior and Gaussian hyper prior, $\mathcal{H}_p$ represents the parameters $\{\alpha, \lambda, \mu_0, \sigma_0, \mu_1, \sigma_1\}$. The hyperparameters can be learned through maximizing the posterior distribution in (6) using some gradient based optimization algorithms. As α and λ control the expectation and variance of the Gamma hyper prior for the inverse length-scales, we can obtain the inverse length-scales $\{\ell_d\}$ with different characteristics through adjusting the settings of α and λ. For example, the difference between the inverse length-scales $\{\ell_d\}$ will be larger if the variance of the Gamma prior is larger.

3 Data Collection and Reconstruction

We collect the students' scores from the department of computer science and technology in a certain university. The data are collected from two grades which are Grade 2010 and Grade 2011. In each grade, there are two classes which are pedagogical class and regular class. In each class, the numbers of students are different which are 47, 51, 23 and 52, respectively. We separate the data from each class into seven groups with each one corresponding to one semester. As different students are likely to choose different courses, the course numbers are different in each group of data. Therefore, we have to employ data-filling methods to reconstruct the data.

When reconstructing the data, the nearest neighbor (NN) is a common data-filling method which is to find the most appropriate data for the missing value. In NN, for the missing score of every course from every student, the score from the most similar student is used to fill the missing score. The similarity is measured by the distance of two students' scores for the chosen courses. Since different students often choose different courses, the simple Euclidean distance are inappropriate for measuring the distance between two samples. For fairness, we compute the averaged distance per course.

4 Experiments

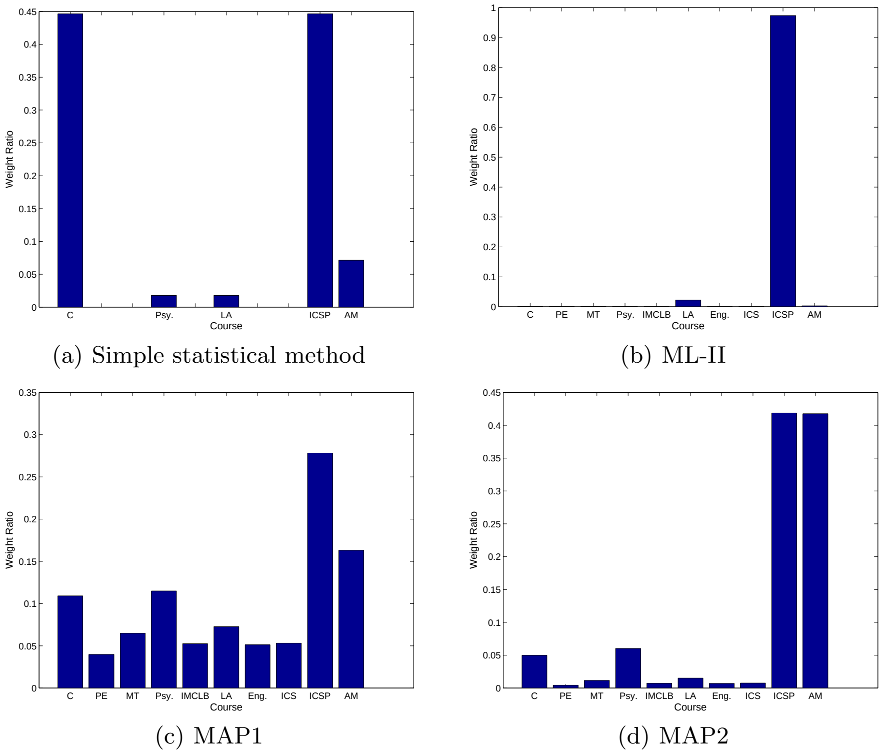
4.1 Experimental Settings

We use two kinds of GPR model selection methods which are ML-II and MAP for learning the importance of courses. The simple statistical method is used as reference. Particularly in the MAP model selection method, we assume two different Gamma hyper prior distributions which have small and large variances respectively for the inverse length-scales. We denote these two MAP methods as MAP1 and MAP2. The Gamma hyper priors are set to $\mathrm{Gamma}(0.1, 0.1)$ and $\mathrm{Gamma}(0.1, 0.02)$. The Gaussian hyper priors for $\log \sigma_f^2$ and $\log \sigma_n^2$ are set to $\mathcal{N}(-1, 1)$ and $\mathcal{N}(-3, 3)$. For both the two GPR model selection methods, the maximum iteration numbers for optimization are set to 1000.

4.2 Experimental Results and Analysis

We demonstrate and analyze our experimental results by taking an example from a certain semester. The analysis of the data from other semesters is similar. Particularly, we analyze the selected courses by different GPR model selection methods as well as the simple statistical method. We plot the details of the key courses from the two classes (pedagogical class in Fig. 1 and regular class in Fig. 2) in the first semester in Grade 2010.

From Fig. 1, we find that the four methods all regard "Introduction to Computer Science and Practice (ICSP)" as the most critical course. This is convictive because "ICSP" contains the basic operations on computer. But beyond that, most results from the simple statistical method only show the appearance instead of the hidden characteristics in the data. For example, it shows that "C" is prone to fail, but it weakens the importance of "Advanced Mathematic (AM)". Differently, MAP1 and MAP2 both put more weights on "AM". ML-II selects the "Linear Algebra (LA)" and "Advanced Mathematic (AM)" as the additional key courses. Through further analysis of the course characteristics, we know that



(a) Simple statistical method

(b) ML-II

(c) MAP1

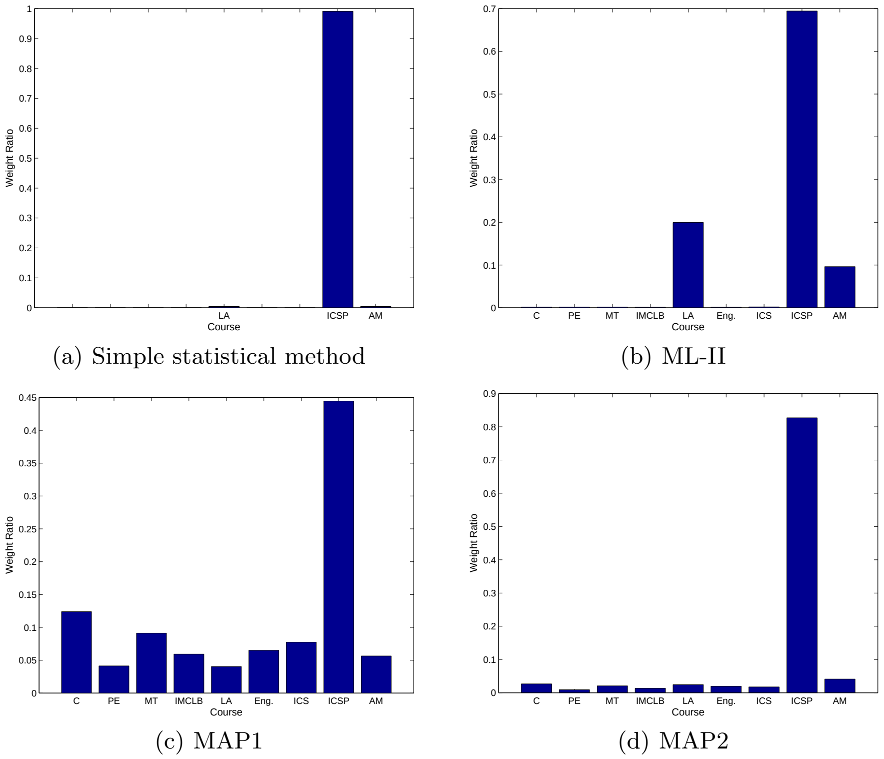
(d) MAP2

Fig. 1. Key course selection results for the pedagogical class in the first semester in Grade 2010.

"AM" is the foundation course for computer related courses. "AM" is recognized as a challenging and important course, especially for pedagogical students who have relative poor mathematical basis. After the analysis of these key courses, it is recommended to focus more on the education of mathematical courses for pedagogical students in this semester.

From Fig. 2, we find that the four methods all regard "ICSP" as the most critical course. This is consistent with the conclusion from the pedagogical class. However, except for "ICSP", the simple statistical method shows that there is few failed courses for regular class. ML-II selects two additional courses, "Linear Algebra (LA)" and "AM ", as key courses while MAP1 and MAP2 treat the other courses almost the same except "ICSP" . Such phenomenon implies that the students in the regular class have better mathematical foundation compared with pedagogical students. It is recommended to balance attentions to every course on the basis of laying emphasis on "ICSP" for regular students in this semester.

Comparing MAP1 and MAP2, we find that the two MAP methods with two different Gamma hyper priors actually obtain two trends of weight ratios. The Gamma hyper prior with larger variance tends to infer the length-scales with



(a) Simple statistical method

(b) ML-II

(c) MAP1

(d) MAP2

Fig. 2. Key course selection results for the regular class in the first semester in Grade 2010.

wide difference. The introduction of hyper priors can overcome the over-fitting to some extent and bring convenience to adjust the magnitude of the difference between the weights.

5 Conclusion and Future Work

We have selected key courses from the academic data by GPR model selection methods. From the experimental results on the real-world data, we conclude that the MAP model selection method based on GPR is a reasonable and flexible method for key course selection. GPR model selection methods can discover the hidden information about courses. Combining the key courses selected by the GPR model selection methods and those selected by the simple statistical method, the ultimately selected courses are very useful no matter for students or educational administrators. From the students' point of view, it can warn the students to pay more attention to some specific courses. From the educational administrators' point of view, it can help to evaluate the students according to the scores of some key courses and then make proper educational policies.

Key course selection is an important task for AEW and it is only at the primary stage of work on AEW. In the future work, we will discuss how to warn students to pay more attention to some specific courses in the next semester by their performance in the current semester.

Acknowledgments

The first two authors Min Yin and Jing Zhao are joint first authors. The corresponding author Shiliang Sun would like to thank support by NSFC Project 61370175.

References

1. Efron, B., Johnstone, I., Hastie, T., Tibshirani, R.: Least angle regression. The Annals of Statistics 32, 407–499 (2002)
2. Ghahramani, Z.: Probabilistic machine learning and artificial intelligence. Nature 521, 452–459 (2015)
3. Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z., Wolff, A.: OU analyse: analysing at-risk students at the open university. Learning Analytics Review, LAK15-1, 1–16 (2015)
4. Rasmussen, C.E., Williams, C.K.I.: Gaussian Process for Machine Learning. MIT Press (2006)
5. Shawe-Taylor, J., Sun, S.: A review of optimization methodologies in support vector machines. Neurocomputing 74, 3609–3618 (2011)
6. Shi, J., Choi, T.: Gaussian Process Regression Analysis for Functional Data. CRC Press (2011)
7. Sun, S.: Infinite mixtures of multivariate Gaussian processes. In: Proceedings of the International Conference on Machine Learning and Cybernetics. pp. 1011–1016 (2013)
8. Sun, S.: Computational education science and ten research directions. Communications of the Chinese Association for Artificial Intelligence 9, 15–16 (2015)
9. Sun, S., Xu, X.: Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction. IEEE Transactions on Intelligent Transportation Systems 12, 466–475 (2011)
10. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58, 267–288 (1996)
11. Titsias, M.K.: Variational learning of inducing variables in sparse Gaussian processes. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. pp. 567–574 (2009)
12. Vidaurre, D., Bielza, C., Larrãnaga, P.: A survey of L1 regression. International Statistical Review 81, 361–387 (2013)
13. Zhang, Q., Hu, X., Zhang, B.: Comparison of L1-norm SVR and sparse coding algorithms for linear regression. IEEE Transactions on Neural Networks and Learning Systems 26, 1828–1833 (2015)
14. Zhu, J., Sun, S.: Single-task and multitask sparse Gaussian processes. In: Proceedings of the International Conference on Machine Learning and Cybernetics. pp. 1033–1038 (2013)