ПОСТРОЕНИЕ МОДЕЛИ КЛАССИФИКАЦИИ ЗЕРНОВЫХ СЫПУЧИХ ВЕЩЕСТВ ПО СПЕКТРАМ ДИФФУЗНОГО ОТРАЖЕНИЯ В БЛИЖНЕЙ ИНФРАКРАСНОЙ ОБЛАСТИ НА ПРИМЕРЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

С. В. Проценко*, В. С. Мишурная, Е. С. Воропай

УДК 543.424.4

Белорусский государственный университет,
220030, Минск, просп. Независимости, 4, Беларусь; e-mail: stas-p0@rambler.ru, voropay@bsu.by

(Поступила 22 октября 2018)

Экспериментально получено подтверждение возможности определения зерновой культуры, имеющей различный помол, по спектру диффузного отражения. В качестве признаков, описывающих спектры диффузного отражения пшеницы и овса различных помола и влажности в ближнем ИК диапазоне, использованы комбинации оптических плотностей и их вторых производных для длин волн 1200, 1422, 1778, 1916 и 2114 нм. На примере логистической регрессии построено 20 моделей классификации по двум признакам: 10 моделей для оптической плотности и 10 моделей для второй производной от оптической плотности, соответствующих выбранным длинам волн. Наилучшие результаты классификации получены с помощью алгоритма, использующего в качестве признаков значения второй производной от оптической плотности на λ = 1778 и 2114 нм.

Ключевые слова: *диффузное отражение, инфракрасная спектроскопия, логистическая регрессия, машинное обучение.*

The possibility of determining a grain corps having a different grinding by the diffuse reflection spectrum was experimentally confirmed. Combinations of optical densities and their second derivatives for the wavelengths of 1200, 1422, 1778, 1916, and 2114 nm were used as features describing the diffuse reflection spectra of wheat and oats of different milling and humidity in the near infrared range. Using the example of logistic regression, 20 classification models based on two features were constructed: 10 models for optical density and 10 models for the second derivative of the optical density corresponding to the selected wavelengths. The best classification results were obtained by an algorithm that used the values of the second derivative of optical density at 1778 and 2114 nm as features.

Keywords: *diffuse reflection, infrared spectroscopy, logistic regression, machine learning.*

Введение. Спектроскопия диффузного отражения сыпучих и порошкообразных веществ в ближнем ИК диапазоне — мощный инструмент количественного анализа, позволяющий быстро и с высокой точностью определять компонентный состав анализируемого материала в лабораторных условиях, а также в условиях технологического процесса со стационарным потоком [1—3]. Причина высокой информативности спектров диффузного отражения сыпучих и порошкообразных веществ в ближнем ИК диапазоне заключается в большом количестве полос поглощения, что позволяет строить более сложные градуировочные уравнения для определения компонентного состава [1]. Требования, предъявляемые к измерению компонентного состава анализируемого материала на лабораторных установках, работающих по принципу приема отраженного излучения, четко стандартизированы в плане предварительной пробоподготовки, а также условий, при которых осуществляются измерения, что обеспечивает высокую повторяемость результатов анализа. Однако при непрерывном изме-

---

BUILDING A MODEL FOR CLASSIFICATION OF GRAIN BULK PRODUCTS BY DIFFUSE REFLECTION SPECTRA IN THE NEAR INFRARED REGION ON THE EXAMPLE OF LOGISTIC REGRESSION

S. V. Protsenko*, V. S. Mishurnaya, E. S. Voropai (*Belarusian State University, 4 Nezavisimosti Prosp., Minsk, 220030, Belarus; e-mail: stas-p0@rambler.ru, voropay@bsu.by*)

рении компонентного состава в технологическом процессе системами, работающими по принципу приема отраженного излучения, условия, при которых осуществляется анализ, изменяются с различной периодичностью, что сказывается на результатах измерений. В этом случае основным источником ошибок является несоответствие установленного градуировочного уравнения условиям, в которых осуществляется анализ, и/или физико-химическим свойствам материала, взятого для градуировки [1, 4—6].

Использование нескольких градуировочных уравнений, соответствующих различным условиям измерения и/или физико-химическим свойствам, не решает данную проблему в полном объеме по причине того, что в условиях работы в технологическом потоке необходим автоматический выбор градуировочного уравнения, максимально соответствующего условиям измерения. Выход из сложившейся ситуации — применение методов машинного обучения, позволяющих распознавать изменение условий по анализу спектров диффузного отражения и автоматически выбирать наиболее подходящее градуировочное уравнение. Поскольку автоматизированное производство подразумевает большое количество измерений с высокой частотой, что эквивалентно большому количеству данных и признаков, описывающих анализируемый материал, использование данной информации позволит учитывать нестационарность потока и/или изменение физико-химических свойств материала, тем самым повысить качество измерений и осуществлять разделение продукции по качеству в соответствии с физико-химическими свойствами [1]. Методы машинного обучения могут применяться для повышения эффективности технологического процесса при производстве продукции в сельском хозяйстве, пищевой промышленности, деревообработке, химической промышленности, фармацевтике путем сбора и анализа получаемых данных. С экономической точки зрения использование алгоритмов машинного обучения позволит проводить гибкую ценовую политику в соответствии с физико-химическими свойствами продукции, а также осуществлять оптимизацию бизнес-процессов.

Методика исследования и пробоподготовки. Для построения модели классификации, учитывающей вариации физико-химических свойств, взято по 16 образцов пшеницы и овса различной влажности и крупности помола зерен. Предварительно образцы пшеницы и овса были помолоты и просеяны через сита с различными размерами ячеек, после чего они увлажнялись, и через двое суток их влажность определялась в сушильном шкафу при температуре 105 °С согласно ГОСТу 13586.5-2015. В табл. 1 представлены характеристики изучаемых зерновых культур по влажности и помолу.

Т а б л и ц а 1. Характеристики образцов пшеницы и овса

Помол, мм	Пшеница				Овес			
	Влажность, %							
5 (без помола)	4.9	7.7	10.6	12.3	6.3	8.8	11.2	14.0
2.5—5	5.1	7.9	10.9	12.7	6.3	8.8	11.2	12.6
1.8—2.5	5.3	8.0	11.0	12.6	6.5	8.6	10.9	13.2
0.8—1.8	5.1	7.3	10.9	12.2	6.6	8.7	11.1	13.3

Спектры диффузного отражения получены на отражательном спектрометре Foss NIRSystems 5000. Каждый образец измерялся дважды, в результате получено 64 спектра для пшеницы и овса. Как видно на рис. 1, *а*, спектры диффузного отражения обладают особенностью — наличием общего наклона и фонового поглощения, что связано со значительным влиянием крупности частиц, составляющих анализируемый материал. Для более крупных частиц характерно увеличение проходимого оптического пути, что дополнительно усиливает поглощение.

Количественный анализ по спектрам диффузного отражения без предварительной математической обработки затруднен в силу наличия общего наклона и фонового поглощения. Для устранения данных эффектов используются различные методы преобразования спектральных данных без потери полезной информации. Один из наиболее популярных методов предварительной обработки — взятие второй производной от оптической плотности D. Как видно из рис. 1, *б*, взятие второй производной позволило исключить влияние общего наклона и фонового поглощения. Эта операция также увеличила информативность через увеличение разрешения полос поглощения.

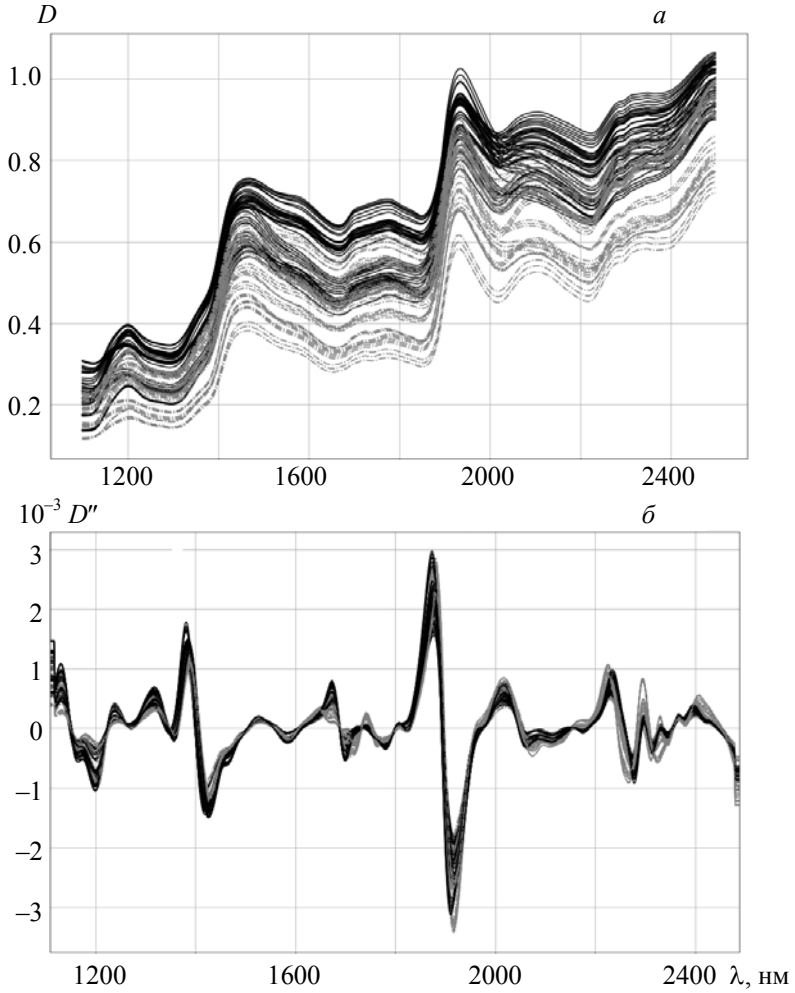

Рис. 1. Спектры диффузного отражения пшеницы (сплошные линии) и овса (штрихпунктир) (*а*) и их вторая производная (*б*)

Алгоритмы машинного обучения: логистическая регрессия. Алгоритмы машинного обучения можно разделить на два класса: обучение с учителем и обучение без учителя. В случае алгоритма, не требующего учителя, классификация осуществляется по определенным правилам, которые устанавливаются в результате анализа признаков, характеризующих объекты выборки. При работе с алгоритмами машинного обучения с учителем каждый объект выборки характеризуется определенным набором признаков и заранее известна принадлежность к тому или иному классу. Для определения параметров модели машинного обучения выборка разделяется на тренировочную и тестовую. Тренировочная используется для определения коэффициентов модели по заданному набору признаков, тестовая предназначена для проверки качества работы алгоритма, осуществляющего классификацию на основании коэффициентов, которые определены на тренировочной выборке [7—9].

Наиболее простым, но в то же время мощным алгоритмом машинного обучения, осуществляющим классификацию, является логистическая регрессия. Данный алгоритм используется как для бинарной, так и для многоклассовой классификации. Логистическая регрессия применяется в таких областях, как медицина, социология, биология, лингвистика и т. д. [7, 8]. Геометрический смысл логистической регрессии заключается в построении разделяющей поверхности в пространстве признаков, описывающих анализируемые объекты, где расстояние от отдельно выбранной точки до разделяющей поверхности есть вероятность принадлежности объекта к тому или иному классу. Для определения этой вероятности используется логистическая функция

$$P(x_1, x_2, ..., x_n) = \{1 + \exp(-[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... \beta_n x_n])\}^{-1}, \quad (1)$$

где x_i — независимые переменные; β_i — коэффициенты логистической регрессии.

Построение модели классификации. Как отмечено выше, для построения модели классификации выборку предварительно необходимо разделить на тестовую и обучающую. Данная процедура выполняется с целью подбора оптимальных коэффициентов логистической регрессии и проверки качества работы алгоритма с использованием метрик качества. Также следует определить признаки объекта, которые будут использованы для обучения алгоритма. В рассматриваемом случае признаками выступают спектральные данные. Для увеличения размера всей выборки каждая пара спектров, соответствующих отдельно взятому образцу, усреднялась. Эта процедура является искусственной, однако она не оказывает негативного воздействия на модель данных, а только способствует увеличению количества признаков. В результате общее количество спектров увеличивается до 96.

В качестве признаков, описывающих пшеницу и овес, выбраны оптическая плотность (D) и ее вторая производная (D'') для следующих длин волн: 1200, 1422, 1778, 1916 и 2114 нм. Выбор объясняется тем, что пшеница и овес испытывают поглощение на рассматриваемых длинах волн.

Процесс поиска наилучшей модели классификации на основе логистической регрессии разделен на два блока: один блок включает рассмотрение моделей, построенных на основании D без предварительной математической обработки, в другом блоке для построения моделей взяты значения D''. В каждом из блоков использовались различные варианты комбинаций признаков, соответствующие D или D'' для длин волн, на которых осуществлялось обучение и тестирование. Качество алгоритмов классификации проверено матрицей ошибок, соотносящей реальные классы и определенные алгоритмом:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

где a_{11} — количество оценок алгоритма (пшеница), совпавшее с истинным классом (пшеница); a_{12} — количество оценок алгоритма (овес), не совпавшее с истинным классом (пшеница); a_{21} — количество оценок алгоритма (пшеница), не совпавшее с истинным классом (овес); a_{22} — количество оценок алгоритма (овес), совпавшее с истинным классом (овес).

Т а б л и ц а 2. Метрики качества алгоритмов классификации-идентификации при работе с оптической плотностью для тестовых и тренировочных выборок

Комбинация длин волн, нм	Тренировочная выборка	Тестовая выборка	Комбинация длин волн, нм	Тренировочная выборка	Тестовая выборка
1200-1422	$\begin{bmatrix} 29 & 7 \\ 7 & 29 \end{bmatrix}$	$\begin{bmatrix} 9 & 3 \\ 2 & 10 \end{bmatrix}$	1422-1916	$\begin{bmatrix} 30 & 6 \\ 9 & 27 \end{bmatrix}$	$\begin{bmatrix} 9 & 3 \\ 2 & 10 \end{bmatrix}$
1200-1778	$\begin{bmatrix} 27 & 9 \\ 7 & 29 \end{bmatrix}$	$\begin{bmatrix} 9 & 3 \\ 2 & 10 \end{bmatrix}$	1422-2114	$\begin{bmatrix} 29 & 7 \\ 7 & 29 \end{bmatrix}$	$\begin{bmatrix} 9 & 3 \\ 2 & 10 \end{bmatrix}$
1200-1916	$\begin{bmatrix} 30 & 6 \\ 8 & 28 \end{bmatrix}$	$\begin{bmatrix} 9 & 3 \\ 1 & 11 \end{bmatrix}$	1778-1916	$\begin{bmatrix} 30 & 6 \\ 10 & 26 \end{bmatrix}$	$\begin{bmatrix} 9 & 3 \\ 2 & 10 \end{bmatrix}$
1200-2114	$\begin{bmatrix} 29 & 7 \\ 8 & 28 \end{bmatrix}$	$\begin{bmatrix} 10 & 2 \\ 5 & 7 \end{bmatrix}$	1778-2114	$\begin{bmatrix} 29 & 7 \\ 8 & 28 \end{bmatrix}$	$\begin{bmatrix} 10 & 2 \\ 5 & 7 \end{bmatrix}$
1422-1778	$\begin{bmatrix} 29 & 7 \\ 7 & 29 \end{bmatrix}$	$\begin{bmatrix} 9 & 3 \\ 2 & 10 \end{bmatrix}$	1916-2114	$\begin{bmatrix} 30 & 6 \\ 9 & 27 \end{bmatrix}$	$\begin{bmatrix} 9 & 3 \\ 2 & 10 \end{bmatrix}$

В табл. 2 представлены матрицы ошибок для моделей классификации-идентификации и соответствующие им комбинации длин волн. Как видно, алгоритм классификации, основанный на анализе D для $\lambda = 1200$ и 1916 нм, показывает наилучшие результаты классификации. Уравнение логистической регрессии для алгоритма, принимающего на вход значения D на $\lambda = 1200$ и 1916 нм (D_{1200}, D_{1916}), имеет вид:

$$P(D_{1200}, D_{1916}) = \{1 + e^{-[0.10 - 0.34D_{1200} - 1.62D_{1916}\}^{-1}}. \quad (2)$$

На рис. 2, *а* представлена разделяющая плоскость, определенная на основании уравнения (2). Значительное количество ошибок обусловлено тем, что спектр диффузного отражения таких слож-

ных по составу объектов, как пшеница или овес, представляет собой результат взаимодействия рядом стоящих полос поглощения, что затрудняет получение информации, соответствующей определенной длине волны.

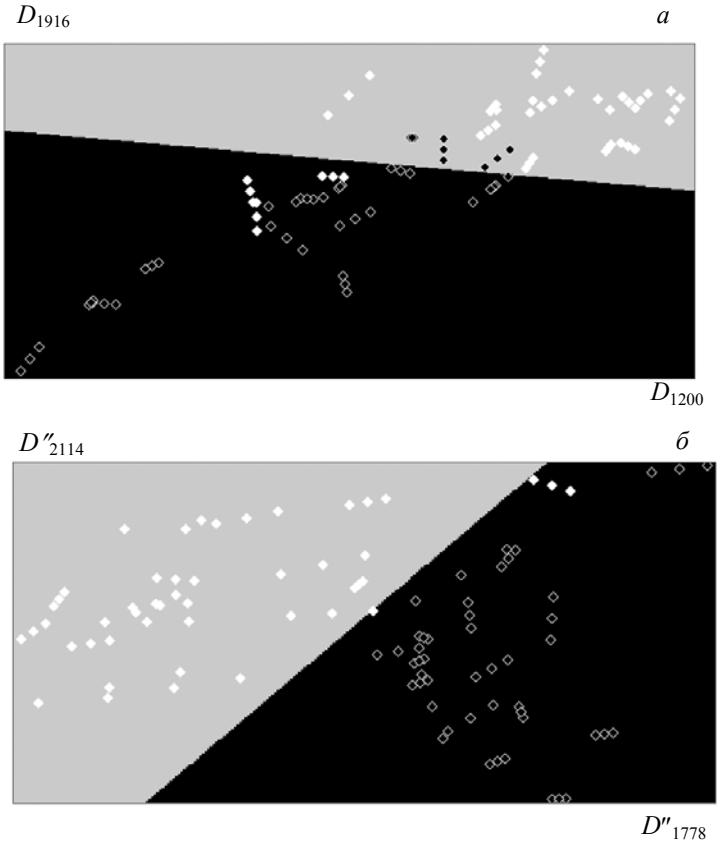

Рис. 2. Разделяющая поверхность, построенная для алгоритма классификации-идентификации пшеницы (♦) и овса (◊): *а* — на основе анализа оптической плотности для λ = 1200 и 1916 нм, *б* — на основе анализа второй производной от оптической плотности для λ = 1778 и 2114 нм

Т а б л и ц а 3. Метрики качества алгоритмов классификации-идентификации при работе со второй производной от оптической плотности для тестовых и тренировочных выборок

Комбинация длин волн, нм	Тренировочная выборка	Тестовая выборка	Комбинация длин волн, нм	Тренировочная выборка	Тестовая выборка
1200-1422	$\begin{bmatrix}32 & 4\\10 & 26\end{bmatrix}$	$\begin{bmatrix}8 & 4\\2 & 10\end{bmatrix}$	1422-1916	$\begin{bmatrix}29 & 7\\8 & 28\end{bmatrix}$	$\begin{bmatrix}7 & 5\\3 & 9\end{bmatrix}$
1200-1778	$\begin{bmatrix}32 & 4\\0 & 36\end{bmatrix}$	$\begin{bmatrix}7 & 5\\0 & 12\end{bmatrix}$	1422-2114	$\begin{bmatrix}31 & 5\\4 & 32\end{bmatrix}$	$\begin{bmatrix}7 & 5\\3 & 9\end{bmatrix}$
1200-1916	$\begin{bmatrix}25 & 11\\11 & 25\end{bmatrix}$	$\begin{bmatrix}8 & 4\\4 & 8\end{bmatrix}$	1778-1916	$\begin{bmatrix}33 & 3\\0 & 36\end{bmatrix}$	$\begin{bmatrix}9 & 3\\0 & 12\end{bmatrix}$
1200-2114	$\begin{bmatrix}30 & 6\\12 & 24\end{bmatrix}$	$\begin{bmatrix}10 & 2\\5 & 7\end{bmatrix}$	1778-2114	$\begin{bmatrix}35 & 1\\0 & 36\end{bmatrix}$	$\begin{bmatrix}9 & 3\\0 & 12\end{bmatrix}$
1422-1778	$\begin{bmatrix}32 & 4\\0 & 36\end{bmatrix}$	$\begin{bmatrix}7 & 5\\0 & 12\end{bmatrix}$	1916-2114	$\begin{bmatrix}31 & 5\\8 & 28\end{bmatrix}$	$\begin{bmatrix}10 & 2\\7 & 5\end{bmatrix}$

Построение алгоритма классификации, работающего со значениями D'', полностью аналогично построению рассмотренного выше алгоритма. В табл. 3 представлены комбинации длин волн, для которых значения D'' использованы при построении алгоритма, совместно с выводом матрицы ошибок, соответствующей тренировочной выборке. Наименьшее количество ошибок дает алгоритм классификации, работающий со значениями D'' на λ = 1778 и 2114 нм. Повышение качества классификации связано с применением второй производной для предварительной математической обработки спектральных данных, что позволяет разрешать рядом стоящие полосы поглощения, а также устраняет общий наклон и фоновое поглощение. Уравнение логистической регрессии для алгоритма, принимающего на вход значения D'' на λ = 1778 и 2114 нм (D''_{1778} , D''_{2114}):

$$P(D''_{1778}, D''_{2114}) = \left\{1 + \exp(-[0.001 + 0.138D''_{1778} - 0.072D''_{2114}])\right\}^{-1}. \quad (3)$$

На рис. 2, *б* представлена разделяющая плоскость, определенная на основании уравнения (3). Полученные результаты подтверждают, что предварительная математическая обработка, заключающаяся во взятии второй производной, способствует повышению качества классификации.

Заключение. Изучены спектры диффузного отражения пшеницы и овса различных помола и влажности в ближнем ИК диапазоне на отражательном спектрометре Foss NIRSystems 5000. Для построения модели классификации на основании логистической регрессии использованы значения оптической плотности для набора длин волн, для которых наблюдается поглощение у пшеницы и овса. С целью минимизация эффекта переобучения использованы только два признака, которые выбраны путем комбинирования всех возможных сочетаний оптической плотности для рассматриваемых длин волн в одном случае и второй производной в другом. На основании анализа матрицы ошибок показано, что алгоритм, работающий со значениями второй производной от оптической плотности на λ = 1778 и 2114 нм, дает наилучшие результаты классификации.

Полученные результаты могут быть использованы для классификации продукции, а также для минимизации ошибки измерения компонентного состава при непрерывном измерении в условиях нестационарного потока. Это позволит проводить более точные измерения системами контроля качества и увеличивать рентабельность предприятий, осуществляющих анализ сыпучей и порошкообразной продукции в потоке.

[1] **В. П. Крищенко.** Ближняя инфракрасная спектроскопия, Москва, Кронн-пресс (1997)
[2] **Е. С. Воропай, В. Г. Белкин, С. В. Проценко, К. В. Говорун, Е. А. Колова.** Вестн. Бел. гос. ун-та, Сер. 1. Физ. мат. информ., № 1 (2016) 16—20
[3] **В. Г. Белкин, С. В. Проценко.** Вестн. Бел. гос. ун-та. Сер. 1. Физ. Мат. Информ., № 3 (2014) 22—25
[4] **С. В. Проценко, Е. С. Воропай, В. Г. Белкин.** Журн. прикл. спектр., **84**, № 6 (2017) 1009—1012 [**S. V. Protsenko, E. S. Voropai, V. G. Belkin.** J. Appl. Spectr., **84** (2017) 1081—1083]
[5] **С. В. Проценко, Е. С. Воропай, В. Г. Белкин.** Материалы междунар. науч.-техн. конф. “Материалы, оборудование и энергосберегающие технологии”, Могилев, 27 апреля 2017 г., БРУ (2017) 43—45
[6] **С. В. Проценко, В. Г. Белкин.** Материалы XXIV междунар. науч.-практ. конф. аспирантов, магистрантов и студентов “Физика конденсированного состояния”, Гродно, 21 апреля 2016 г., ГрГУ (2016) 174—176
[7] **D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant.** Applied Logistic Regression, **398**, John Wiley & Sons (2013)
[8] **S. Menard.** Applied Logistic Regression Analysis, **106**, Sage (2002)
[9] **E. Alpaydin.** Introduction to Machine Learning, MIT press (2009)