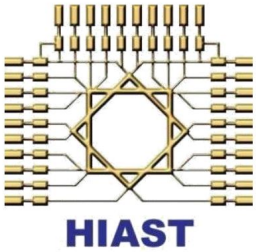
الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا
قسم المعلوميات
العام الدراسي 2017/2018



تقييم جودة نصوص اللغة الانكليزية

مشروع السنة الرابعة

إعداد
فاروق حجابو

إشراف
د. غيداء ربداوي م. رياض سنبل

2 آب 2018

الملخص

المحتويات

قائمة الأشكال

قائمة الجداول

الاختصارات

SVM Support Vector Machines

المصطلحات

Artificial Intelligence	الذكاء الصنعي
Machine Learning	تعلّم الآلة
Natural Language Processing	معالجة اللغات الطبيعية
Supervised Learning	التعلم تحت الإشراف
Unsupervised Learning	التعلم بدون إشراف
Semi-Supervised Learning	التعلم نصف المشرف عليه
Reinforcement Learning	التعلم بالتعزيز
Classification	التصنيف
Regression	الانحدار
Training Set	معطيات التدريب
Training Instance	مثال تدريبي
Accuracy	الدقّة
Clustering	التجميع
Features	ميزات
Feature Extraction	استخراج الميزات

الفصل الأول

التعريف بالمشروع

يُمهّد هذا الفصل للمشروع، حيث يُبيّن فكرة المشروع وأهميتها والأهداف المرجوّة منه. ويذكر المتطلبات الوظيفية وغير الوظيفية للمشروع.

1.1 مقدمة

الفصل الثاني

الدراسة المرجعية

يبيّن هذا الفصل الدراسة المرجعية للمشروع. يبدأ بتقديم مفاهيم تعلّم الآلة ومراحلها المختلفة والمعايير المعتمدة لتقييمها. ويقدّم مفاهيم ومراحل معالجة اللغات الطبيعية. وأخيراً يسرد بعض الأوراق الأبحاث العلمية المتعلقة بهذا المشروع، ويوضح المنهجيات المتبعة فيها.

1.2 تعلم الآلة

تعلم الآلة Machine Learning هو فرع جزئي من الذكاء الصنعي Artificial Intelligence. يُقصد بتعلم الآلة مجموعة الأدوات والمفاهيم والمنهجيات المستخدمة لبرمجة الحواسيب بطريقة تسمح لهذه الحواسيب بالتعلم من المعطيات [11].

ويمكن أيضاً تعريفه بشكل أكثر عمومية كالتالي:

"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed."
–Arthur Samuel, 1959

كما يعتبر التعريف التالي تقني وأكثر دقة:

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."
–Tom Mitchell, 1997

على سبيل المثال، النظام الذي يقوم بفلترة الإيميلات إلى إيميلات مؤذية spam وإيميلات غير مؤذية non-spam، يستخدم منهجيات تعلم الآلة. يقوم هذا النظام بتعلم طريقة التمييز بين هذين النوعين من الإيميلات باستخدام عدد كبير من الأمثلة والمعطيات المصنفة مسبقاً. نسمي هذه المجموعة من الأمثلة بمعطيات التدريب Training Set، وكل مثال منها نسميه مثال تدريبي Training Instance.

في هذه الحالة، المهمة T هي تصنيف الإيميلات الجديدة إلى إيميلات مؤذية وإيميلات غير مؤذية، الخبرة E هي مجموعة معطيات التدريب، ومؤشر قياس الأداء P يمكن تعريفه بعدّة طرق؛ فمثلاً يمكننا استخدام نسبة عدد الإيميلات التي تم تصنيفها بشكل صحيح إلى عدد الإيميلات الكلي (هذا المعيار يسمى الدقّة Accuracy كم سنرى لاحقاً).

1.1.2 تصنيفات تعلم الآلة

يمكن تصنيف أنظمة تعلم الآلة وفق عدّة معايير. التصنيف الأكثر شهرة يعتمد على آلية التدريب، وهو كالتالي:

- التعلم تحت الإشراف Supervised Learning: وهي حالة أن تكون الأمثلة التدريبية متوفرة مع الخرج label المرتبط بها. وهذه حالة مثال تصنيف الإيميلات المطروح سابقاً. حيث أن معطيات التدريب هي مجموعة كبيرة من الإيميلات المصنفة مسبقاً من قبل البشر إلى إيميلات مؤذية وإيميلات غير مؤذية.
- التعلم بدون إشراف Unsupervised Learning: وهي حالة أن تكون معطيات التدريب موجودة ولكنها غير مصنفة unlabeled أو غير مرتبطة بخرج معيّن. على سبيل المثال، قد ترغب شركة في تصنيف زبائنها إلى عدّة مستويات، زبائن من الدرجة الأولى، زبائن من الدرجة الثانية، وهكذا. فيمكن استخدام تعلم الآلة لاكتشاف بعض الأنماط الموجودة في معطيات الزبائن واكتشاف هكذا تصنيف. وهذا ما يُعرف بالتجميع Clustering.
- التعلم نصف المشرف عليه Semi-Supervised Learning: وهي حالة وسيطة بين التصنيفين السابقين. تكون فيها بعض أمثلة التدريب مرتبطة بخرج معيّن (غالباً تشكل النسبة الصغيرة)، وتكون باقي الأمثلة غير مرتبطة بخرج. تنطبق هذه الحالة على مثال تصنيف الإيميلات في حال لم تكن جميع معطيات التدريب مصنفة بشكل مسبق.
- التعلم بالتعزيز Reinforcement Learning: وهي الحالة التي يتخاطب فيها النظام مع بيئة أخرى. تقدم له هذه البيئة نتائج feedback بناءً على أفعاله. هذا الصنف ينطبق على الخوارزميات المستخدمة لتدريب الأنظمة

التي تتعلم الألعاب. حيث يقوم النظام بمجموعة من الأفعال actions ضمن بيئة اللعبة، وبناءً على النتائج (تحسّن نتيجته أو انخفاضها) يغيّر أفعاله اللاحقة.

وعلى وجه الخصوص يمكن تصنيف التعلم تحت الإشراف بحسب نوع الخرج المرتبط بمعطيات التدريب. تصنّف بشكل أساسي عريض كالتالي:

- التصنيف Classification: يكون الخرج المرتبط بكل مثال تدريبي هو صف class محدد من مجموعة صفوف. عدد هذه الصفوف قد يكون 2، 3، إلخ. في مثال تصنيف الإيميلات السابق، عدد الصفوف هو 2، حيث أن كل مثال تدريبي (إيميل معيّن من معطيات التدريب) هو إمّا مؤذي أو غير مؤذي.
- الانحدار Regression: يكون الخرج المرتبط بكل مثال تدريبي هو عدد حقيقي. مثل مسألة التنبؤ بسعر منزل بمعرفة معلومات عنه مثل مساحته، عدد الغرف، إلخ.

2.1.2 المراحل اللازمة لتطبيق تعلم الآلة

إذا عدنا إلى مثال تصنيف الإيميلات، حيث قلنا أن معطيات التدريب هي مجموعة من الإيميلات المصنفة بشكل مسبق إلى إيميلات مؤذية وإيميلات غير مؤذية. يمكن أن نسأل هنا: ما هو تحديداً الدخل؟ أي كيف سنعبّر عن الإيميل؟ بالطبع يمكن اعتبار الإيميل كنص؛ فهو مجموعة من الكلمات والرموز. ولكن كما سنرى لاحقاً، من الصعب على معظم خوارزميات تعلم الآلة التعامل مع نص خام. ولذلك هناك مرحلة تسبق مرحلة تنفيذ خوارزميات تعلم الآلة وهي مرحلة تحويل النص إلى ما يسمى بالميزات Features.

فمثلاً يمكن أن نعبّر عن نص الإيميل بميزاته، مثل عدد الكلمات، عدد الجمل، تواتر وجود كلمات مفتاحية محددة، إلخ. نلاحظ الآن في هذه الحالة أننا نتعامل مع الإيميل كشعاع من الميزات feature vector وهذا أمر مناسب جداً للعديد من خوارزميات تعلم الآلة. أيضاً إن الميزات التي ذكرناها هي ميزات عددية numerical features، ولكن بشكل عام يمكن أن تكون الميزات هي ميزات نصية string features أو ميزات صنفية categorical features، إلخ. ويمكن أيضاً تنميط الميزات بشكل مختلف أو أكثر دقة مثل تصنيف الميزات العددية إلى ميزات مستمرة continuous features وميزات متقطعة discrete features. وتعود طريقة التنميط إلى التطبيق أو خوارزميات تعلم الآلة المستخدمة. تسمى هذه المرحلة بمرحلة استخراج الميزات Feature Extraction.

في التطبيقات الواقعية تسبق المرحلة السابقة مرحلتين أساسيتين. مرحلة تجميع المعطيات، ومرحلة تنظيفها. تتم عملية تجميع المعطيات بحسب التطبيق. فمثلاً قد تكون المعطيات هي نتيجة استبيانات، أو إحصائيات، أو تم الحصول عليها من مواقع إلكترونية، إلخ. مرحلة تنظيف المعطيات تهدف إلى التأكد سلامة المعطيات قبل استخدامها. وقد تتم هذه العملية بشكل يدوي أو بشكل مؤتمت وذلك بحسب مصدر المعطيات ونظافتها.

3.1.2 خوارزميات تعلم الآلة

كما رأينا في الفقرة 1.1.2، هناك العديد من أصناف المسائل الممكن حلها باستخدام تعلم الآلة. تصنف خوارزميات تعلم الآلة تبعاً لصنف المسألة التي تقوم بحلها. فمثلاً يمكن استخدام الانحدار الخطي Linear Regression [11] لحل مسائل الانحدار. أو استخدام خوارزمية K-Means Clustering [11] لحل مسائل التجميع.

سنمهد في هذه الفقرة لأهم الخوارزميات المستخدمة في هذا المشروع. وهي: شجرة القرار Support Vector Machines (SVM) وجب التنويه إلى أنه تم النظر إلى المسألة المطروحة في هذا المشروع على أنها مسألة تصنيف. لمزيد من التفاصيل انظر إلى الفقرة

J48

المراجع

[1] Sarah E Schwarm and Mari Ostendorf. "Reading level assessment using support vector machines and statistical language models". In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. 2005, pp. 523–530.

[2] Rong Zheng et al. "A framework for authorship identification of online messages: Writing-style features and classification techniques". In: Journal of the American society for information science and technology 57.3 (2006), pp. 378–393.

[3] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. "Cognitively motivated features for readability assessment". In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. 2009, pp. 229–237.

[4] Sarah E Petersen and Mari Ostendorf. "A machine learning approach to reading level assessment". In: Computer speech & language 23.1 (2009), pp. 89–106.

[5] Xiaofei Lu. "Automatic analysis of syntactic complexity in second language writing". In: International journal of corpus linguistics 15.4 (2010), pp. 474–496.

[6] Sowmya Vajjala and Detmar Meurers. "On improving the accuracy of readability classification using insights from second language acquisition". In: Proceedings of the seventh workshop on building educa-

tional applications using NLP. Association for Computational Linguistics. 2012, pp. 163–173.

[7] Sowmya Vajjala and Detmar Meurers. "Readability assessment for text simplification: From analysing documents to identifying sentential simplifications". In: ITL-International Journal of Applied Linguistics 165.2 (2014), pp. 194–222.

[8] Sowmya Vajjala. "Analyzing text complexity and text simplification: connecting linguistics, processing and educational applications". PhD thesis. Ph. D. thesis, University of Tübingen, 2015.

[9] Sowmya Vajjala and Detmar Meurers. "Readability-based sentence ranking for evaluating text simplification". In: arXiv preprint arXiv:1603.06009 (2016).

[10] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. "Text readability assessment for second language learners". In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. 2016, pp. 12–22.

[11] Aurélien Géron. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.", 2017.

[12] Riad Sonbol. Extracting Business Process Models from Natural Language Texts. Higher Institute for Applied Sciences and Technology, 2017.

[13] Sowmya Vajjala and Ivana Lucic. "OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification". In: (2018).