Peer-graded Assignment: Machine Learning Project Plan

You passed!
Congratulations. You earned 23.5 / 27 points. Review the feedback below and continue the course when you are ready. You can also help more peers by reviewing their submissions.

Review assignments

Instructions | My submission | Discussions

An automated system for disease detection

Submitted on August 9, 2020

Shareable Link

PROMPT

Provide a brief description of the company and project that you will be proposing. Remember, you can copy details from your previous assignment, but make sure you provide an overview of the business, stage of readiness, and business objectives.

We are going to propose a project on creating an automated system that can detect the diseases, such as diabetes, Alzheimer, cancer, melanoma in the early phase of the disease development. Considering that these diseases are very dangerous to person life, I would like to develop a system that can identify the existence of the disease as fast as possible, which can speed up drug development and medical treatment, consequently leading to complete cure from the disease. Currently, the detection of the aforementioned diseases is done by medical doctors. However, considering busy schedule of the doctors and large amount of pay for the consultation with them, significantly large number of individuals cannot afford to benefit from early detection of these hazardous illnesses. Therefore, the system of early detection of the diseases can be automated using the power of Machine Learning, and more specifically Deep Learning. Recently, Convolutional Neural Networks (CNNs) based autoencoding models have advanced the field of image segmentation. Hence, we are going to benefit from the CNN based binary image segmentation models to create the proposed automated system for an early detection of disease.

The business outcome of the project is to benefit from highly qualified, accurate and fast detection of the diseases. After training the Deep Learning model, it will be uploaded on the cloud, so that the patients can upload their images and obtain the result about the existence or absence the disease of their interest. After acquiring the results, the consumers will be asked to fill a short questionnaire on the experience of the using the system. The business success will be evaluated by the satisfaction of the consumers of the automated system. The evaluation will also use the power of Machine Learning. We are going to use online learning for this purpose, meaning that the learning process will never finish and the questionnaire data will increase as the consumers number grow.

An automated system will be detecting the illnesses, such as diabetes, Alzheimer, cancer, melanoma in the early phase of the disease development. Considering that, these diseases are very dangerous to person life, we would like to develop a system that can identify the existence of the disease as fast as possible, which can speed up drug development and medical treatment, consequently leading to complete cure from the disease. Also, by doing research on this field, we would like to assist people who have no access to high-quality medical services due to lack of financial resources. Since, the automated disease detection system will be accessible to everyone who has the internet connection and very basic computer skills. People should just upload the medical images on the system and obtain the information about the existence or absence of a particular disease.

Regarding risks of the proposed system, the primary problem is to obtain data for initial training of the image segmentation model. Since we are going to create the model in the medical area, it will be hard to acquire the desirable amount of data. There are several reasons for this problem. First, as image segmentation models are the representative of supervised learning, the training data must be labelled. However, considering the busy schedule of doctors and tedious process of labelling, it may be difficult to cooperate with the doctors on labelling the medical images. This issue can be partially addressed by crowdsourcing, that is by hiring non-expert people to label the data. However, it can lead to the second problem, which is low-quality training data. Considering that the medical applications require very high degree of accuracy, this problem can be crucial risk to the business. Third, even if doctors may agree on labelling the data for image segmentation, the amount of the data might be small. Since DL methods require millions of training images to output accurate segmentation, this problem also may cause significant amount of risk to the business.

RUBRIC

Company overview

- 0 points
  I have no idea what this company does or how the project relates to any measurable goals.
- **1 point**
  There is a clear description of the company and what product or service it offers.

Business objectives

- 0 points
  I have no idea what this project is or how it is supposed to support business objectives.
- **1 point**
  There is a description of a business outcome and related project

Stage of Machine Learning Readiness

- **~~0 points~~ +0.5 pts because of a tie**
  No machine learning readiness stage is given.
- 1 point
  Yes, the business has been positioned in the machine learning readiness stages.

PROMPT

Who are the major stakeholders for this project? Describe in terms of roles rather than actual names, if you are using a real-life example. Remember to think in terms of users, developers, and business interests and explain why these people have a stake in the project's success.

The stakeholders of the project is the Ministry of Healthcare, the city office, and the Association of Doctors of our city.

RUBRIC

Users identified

- 0 points
  The users of the final project have not been described in any way.
- **~~1 point~~ +1.5 pts because of a tie**
  A description of the users has been provided but is incomplete, either because no explanation was provided or because a significant type of user was missed.
- 2 points
  A clear description of all the relevant users has been provided.

Developers identified

- 0 points
  The people involved in project development have not been described.
- **~~1 point~~ +1.5 pts because of a tie**
  Some participants in project development have been identified but there are significant gaps or the supporting explanation is unclear.
- 2 points
  There is a full description of all relevant stakeholders in the development of the project. The explanation is concise and clear.

Business interests identified

- 0 points
  No business parties were identified or no explanation provided for the relevance of the business role to the project
- **~~1 point~~ +1.5 pts because of a tie**
  Some business parties were identified but significant stakeholders were missed, or the explanation for their relevance was incomplete.
- 2 points
  A complete and clear case was made for all of the relevant business stakeholders.

PROMPT

What machine learning QuAM or model are you building to support the business outcome you described? Describe the question it's answering and how you will be evaluating its performance. Be sure to explain **how** it supports the business outcome.

An automated system will be detecting the illnesses, such as diabetes, Alzheimer, cancer, melanoma in the early phase of the disease development. Considering that, these diseases are very dangerous to person life, we would like to develop a system that can identify the existence of the disease as fast as possible, which can speed up drug development and medical treatment, consequently leading to complete cure from the disease. Also, by doing research on this field, we would like to assist people who have no access to high-quality medical services due to lack of financial resources. Since, the automated disease detection system will be accessible to everyone who has the internet connection and very basic computer skills. People should just upload the medical images on the system and obtain the information about the existence or absence of a particular disease.

RUBRIC

Machine learning question identification

- 0 points
  No actual question was identified
- **~~1 point~~ +1.5 pts because of a tie**
  A question was identified but it's not totally clear how a QuAM or machine learning model could be built to answer that question.
- 2 points
  There is a well-formed question that presents a clear target for the machine learning algorithm. The purpose of the model or QuAM is clear and defensible as a machine learning project.

Appropriate evaluation

- 0 points
  No evaluation metrics were described or the evaluation metrics that are given do not make sense in relation to the machine learning question.
- **~~1 point~~ +1.5 pts because of a tie**
  A valid evaluation metric is given but it is not clear how it relates to the machine learning question or no explanation is given for why that metric is a good one.
- 2 points
  A clear case for a specific evaluation metric is given. The metric makes sense for the context of the problem.

Relation to business objectives

- 0 points
  The proposed QuAM does not seem to relate to the target business objectives.
- **~~1 point~~ +1.5 pts because of a tie**
  The proposed QuAM may relate to the business objectives described above but the explanation is inaccurate and/or poorly supported. The proposed evaluation does not seem relevant for the stated business objectives
- 2 points
  The proposed QuAM clearly relates to the target business objectives and the explanation is clear. The evaluation metric is relevant for the business objectives.

PROMPT

What data will be using to build this project? Where will the labelled training data come from? What kind of data is it, before and after processing? How will the raw data be transformed for both learning and using the model?

We are going to use data from the different hospitals of our city. The data for training our Deep Learning model will be microscopy and dermoscopy images taken in the hospitals. Certainly, the data from different sources will need pre-processing, since the system of data collection varies from hospital to hospital. For example, the medical imaging techniques might be different in each hospital, which may result in different size of the images in the dataset. Also, considering the large size of the images, we may want to reduce its dimensions, using ML techniques such as principal component analysis, which will create fewer features by retaining the desired variance of the data. After these pre-processing steps, the data will be ready to be inputted into Deep Constitutional Neural Network in order to achieve the state-of-the-art performance.

RUBRIC

Data outline

- 0 points
  It is not actually clear what data will be used. If details are provided, they're so vague as to be impossible to speculate about.
- 1 point
  A reasonable description of the data has been given. However, it is not clear either what form the raw data takes or what data will actually be used for training. The description of the raw data sources and type or learning data format and type is incomplete.
- **2 points**
  A complete and clear description of the data is provided. It is clear what the raw data is and where it comes from. The structure of the learning data is clearly described and makes sense.

Data appropriateness

- 0 points
  The data is not clearly described, or does not seem to relate to the problem at hand, or is impossible to get into appropriate format.
- 1 point
  The data seems related to the given machine learning problem but it is not clear how it will actually be transformed appropriately, or no case is made for why the data is relevant for the prediction task.
- **2 points**
  The structure of the data is clearly explained and appropriate for the task, both for the raw data sources and for the learning data.

Labels (or ground truth data)

- 0 points
  There is no description of the actual labels on the data, or there is a description but it is unclear how such labels could be acquired.
- 1 point
  Some description of the appropriate labels is given, but the labels seems inappropriate for the task or difficult to acquire or validate.
- **2 points**
  The process for acquiring labeled data is clearly described and plausible. The labels make an appropriate target for the QuAM.

PROMPT

Communication is important. Identify one of the stakeholders described above and design an experiment to convince them of the validity of your QuAM. Be sure to explain what metric you will use and how you will argue its appropriateness. It can help to think in terms of what graph will illustrate your point.

Because the majority of the stakeholders of the proposed system are doctors, it will be easy to communicate with them in terms of medicine related issues. As every project needs a domain expert during the data processing phase of the project, our company may benefit a lot from the stakeholders knowledge in the area of their expertise. However, it may be hard for the stakeholders to understand the details related to the QuAM building process. Therefore, we are not going deep into details about the model, but attempt to explain the QuAM building at only high-level. Regarding the evaluation of the QuAM, we are going to use the most popular metrics for the image segmentation, which are intersection over union (IoU), and Dice coefficient. Although, these metrics might look mysterious for the stakeholders, we will explain them the calculation process in very detailed way and obtain feedback from them about the validity of the evaluation metrics.

RUBRIC

Metric makes sense

- 0 points
  No metric was given or it's not clear how it can actually be measured, or the experiment described will not actually explain the results as needed.
- 1 point
  The given metric is measurable but it's not totally clear how it will be measured or why it supports the validity of the QuAM.
- **2 points**
  I understand exactly what the experiment is meant to show and how it will do so. It will definitely provide guidance on the QuAM's performance.

Experiment is meaningful to stakeholder

- 0 points
  The stakeholder was not identified, the experiment is unclear, or the connection between the stakeholder's goals and the actual experiment is unclear.
- 1 point
  The experiment will probably make sense to the stakeholder but it's not obvious or there is a much better way to make the case.
- **2 points**
  The experiment described will clearly matter to the stakeholder and is an excellent method for communicating the model's success to this person.

Communication of metric will be clear to stakeholder

- 0 points
  No plan for communicating the metric was given or the plan given does not seem to actually communicate what it should.
- 1 point
  The appropriate things are being communicated but there are better ways to do so, or the communication method is poorly explained.
- **2 points**
  The visualization or comparison of the metric clearly communicates what it should in a way that the stakeholder will understand.

Edit submission

Comments

Comments left for the learner are visible only to that learner and the person who left the comment.

BO

Share your thoughts...