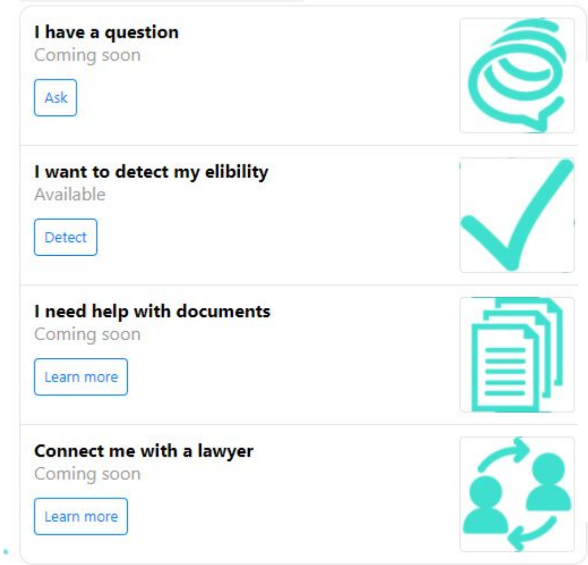
Destin AI, Machine Learning, Assignment 1

Machine Learning and Your Venture

1. How does your team intend to use machine learning?

We are going to be using various Machine Learning Tools which are elaborated on below mentioned sections. Overall, we will use Natural Language Processing to build up chatbot so that immigrants can get answers to their questions, check their eligibility and platform so that people can prepare their documents on their own without a need of the lawyer or simply connect with the lawyer and make sure that their case is well taken care of. Below is the snapshot from our Chatbot and Platform.

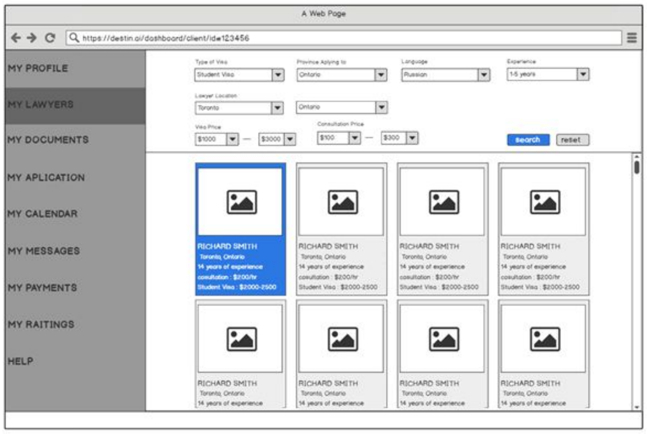


I have a question
Coming soon

I want to detect my elibility
Available

I need help with documents
Coming soon

Connect me with a lawyer
Coming soon



A Web Page

https://destin.ai/dashboard/client/id=123456

Type of Visa: Student Visa
Province Aplying to: Ontario
Language: Russian
Experience: 1-5 years
Lawyer Location: Toronto, Ontario
Visa Price: $1000 — $3000
Consultation Price: $100 — $300

RICHARD SMITH
Toronto, Ontario
14 years of experience
cosultation : $200/hr
Student Visa : $2000-2500

RICHARD SMITH
Toronto, Ontario
14 years of experience
cosultation : $200/hr
Student Visa : $2000-2500

RICHARD SMITH
Toronto, Ontario
14 years of experience
cosultation : $200/hr
Student Visa : $2000-2500

RICHARD SMITH
Toronto, Ontario
14 years of experience
cosultation : $200/hr
Student Visa : $2000-2500

RICHARD SMITH
Toronto, Ontario
14 years of experience

RICHARD SMITH
Toronto, Ontario
14 years of experience

RICHARD SMITH
Toronto, Ontario
14 years of experience

RICHARD SMITH
Toronto, Ontario
14 years of experience

(a) What are the inputs? (e.g. images, videos, text, genomic sequences). Be as specific as possible.

- *Text* - for below purposes:
    - Questions asked by the immigrants
    - Immigrant/Lawyer Application profile data

- *Images - for below purposes:*
    - Extracting information from passports and other application documents

(b) What are the outputs? (e.g. predictions, decisions, visualizations). Be as specific as possible.

Prediction

- To calculate probability of applicants' eligibility for the specific visa programs

Decision

- Provide answers to our clients' immigration related questions
- Recommending suitable immigration lawyers based on the clients' preferences

(c) What techniques are you using/do you plan to use? is could be as specific as classes of techniques (e.g. computer vision, NLP) or even algorithms (e.g. recurrent neural networks).

- Natural Language Processing - Processing of immigration questions and answers
- Neural Networks - Prediction algorithms based on client cases (RNN/LSTM/CNN)
- Clustering - Processing of similar semantically related sentences
- Computer Vision - Extracting features from document images

(d) How do you plan on accessing or implementing such techniques? Be specific about any software frameworks or machine learning APIs that you currently use or plan to use.

We plan to leverage the open source libraries to implement above mentioned techniques such as:

- TensorFlow - Open source software library for implementing Neural Networks
- Theano - Open Source Machine Learning API
- Gensim library: To create Word2Vec/Doc2Vec vectors
- Skipthought models: Techniques to find semantically related sentences.

(e) What are your intended computational needs over the next 6 months? You may wish to discuss cloud compute resources, and/or the use of hardware accelerators.

We will definitely need compute resources for training our models and we intend to use either Google Compute Engine or AWS Cloud.

2. What are your team's data needs or assets?

(a) Discuss the data sources (internal or external) that your team intends to exploit.

We intend to exploit data from all kinds of sources to make our prediction algorithms perform much better. Here are some of them:

- CIC Help Centre http://www.cic.gc.ca/english/helpcentre/index-featured-can.asp
- Facebook Groups regarding immigration where people ask questions
- Our current clients' most frequently asked questions
- Periodic interaction of lawyers and immigrants within our chatbot and platform

(b) Will your team have access to any proprietary data assets or acquisition methods?

Not yet - because we will organically collect the data ourselves as we are dealing with sensitive legal data.

3. Tell us about the machine learning experience on your team.

(a) Please describe any formal (e.g. classroom) or informal (e.g. MOOC) your team members have taken. Please be specific about who has received the training, from whom, and when.

- Mahammad Ismayil - AI course at Waterloo University, Machine Learning by Andrew Ng, Neural Networks by Geoffrey Hinton and Deep Learning by Andrew Ng on Coursera
- Yangqi Xu - Neural Networks and Deep Learning by Andrew Ng on Coursera, Introduction to Artificial Intelligence by Microsoft on eDX

(b) Are there any topics within AI or machine learning that would be helpful to your team if covered in the NextAI curriculum?

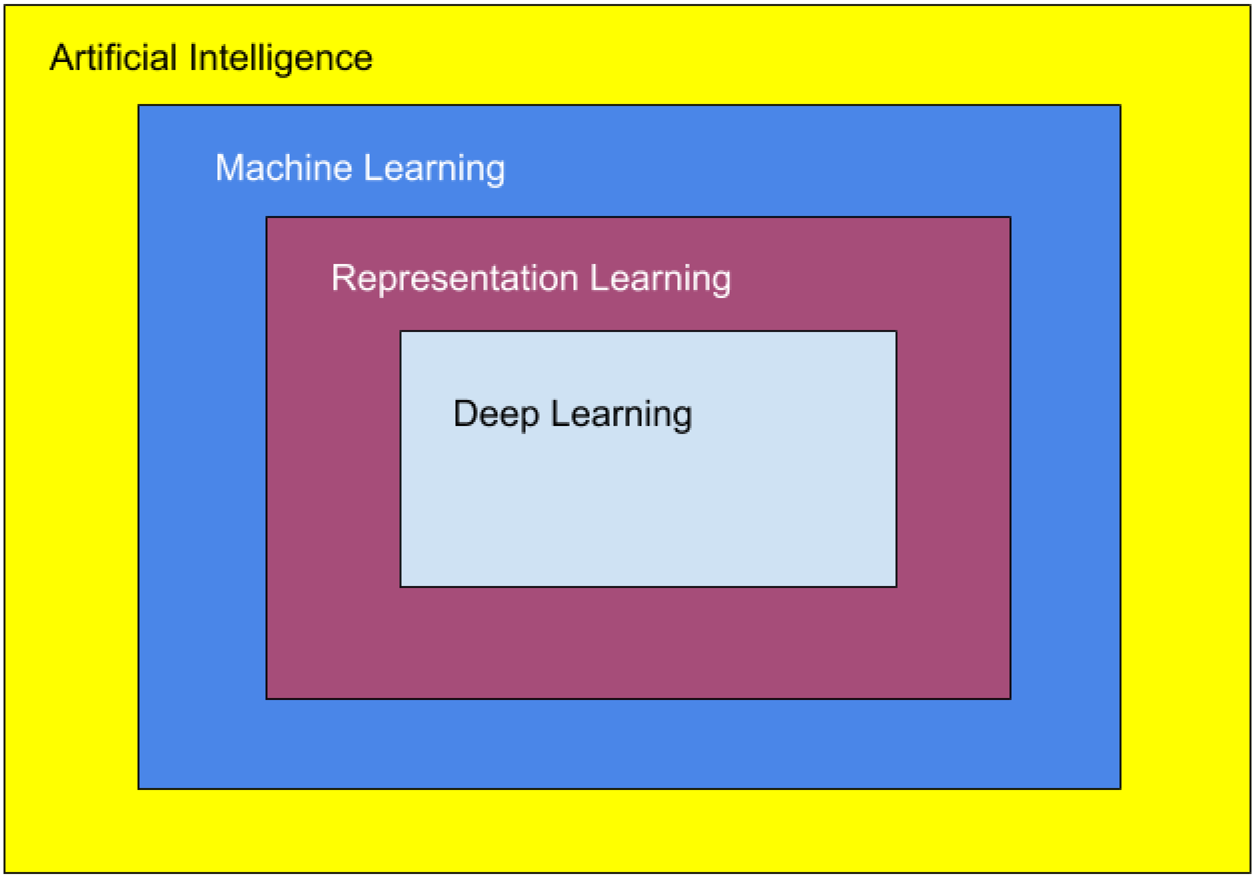
It would be really helpful to cover the following topics in the curriculum:

- Natural Language Processing
- Convolutional Neural Network
- RNN and LSTM
- Autoencoders and PCA

Technical Background

Question 1:

a)



b) 1. Increase in accuracy and complexity of the algorithms. Pre 2006, there were issues with backpropagation algorithm in terms of extending the neural network beyond 2-3 hidden layers. It was not making a better use of multiple hidden layers resulting in a situation where algorithm would get stuck in local optima. This issue was related to the initiation weights in a neural network. In recent years, use of layer-wise greedy unsupervised learning techniques solved this problem. As a result, object detection and speech recognition algorithms have improved dramatically. On the other side, increasing complexity has resulted in revolutionary algorithms such as reinforcement learning, LSTM and neural Turing machines.

2. Tremendous Increase in computational power. With the advent of faster, distributed general purpose CPU and GPU systems, faster network connectivity and better software infrastructure, Neural networks can be trained faster. Today we have enough computational resources to run much larger models. According to Geoffrey Hinton all the techniques that did not work in 1986 worked in 2006. This and the development of new theories have led to major improvements.

3. Increase in availability of labeled data. Backpropagation depends on labelled data to adjust weights in multiple layers. This increase has been driven by increasing digitization of society. We interact with computers more than ever and all these computers are interconnected to each other which makes it easier to collect and centralize data to be used as a huge dataset for machine learning algorithms.

Question 2:

a) Derivative of a function $f(x)$ at point x is denoted as $f'(x)$ or $\frac{dy}{dx}$. It is equal to the slope of that function at that point. In other words, it tells us how the output changes with respect to a small change in input.
Partial derivative of a function $f(x_1, x_2, ..., x_n)$ at variable x_i is denoted as $\frac{\partial}{\partial x_i} f(x_1, x_2, ..., x_n)$. As opposed to whole derivative, it measures how output changes when only one variable x_i changes.
Gradient on the other side, generalizes this notion of derivative by vectorizing all the partial derivatives in one vector and is denoted as $\nabla_x f(x)$ where x is a vector of all the arguments and result is a vector of all partial derivatives.

b) The goal of the machine learning algorithms is to minimize the cost or error function which translates to finding the point where this function obtains its lowest value. In order to find this point, we have to move in a direction of decreasing function values. This is where gradient helps us to move the points in that direction.

Question 3:

a) It speeds up the computation process by leveraging the parallelization of operations.

b) The reason is that squared L^2 norm is more convenient to work with mathematically and computationally. Each derivative of the squared L^2 norm with respect to each element of x has a dependency only on the corresponding element of x, whereas derivatives of L^2 norm depend on the entire vector of x.

c) Diagonal matrices allow us to derive some general, but yet less expensive machine learning algorithms in many cases. They are computationally efficient when it comes to multiplying matrices.

d) Symmetric matrices often arises when some function of two arguments is used to generate matrix entries that does not depend on the order of the arguments. For example, if *A* is a matrix of distance measurements, with A(i, j) giving the distance from point i to j, then A(i,j)=A(j,i), because distance functions are symmetric.

Question 4:

a) The three most common types of Gaussian covariance matrices are precision, diagonal and isotropic. Precision matrix are the inverse of covariance matrix. Diagonal covariance matrix is a covariance matrix where only diagonal elements are non-zero. Isotropic matrix on the other side is a scalar times the identity matrix.

b) We usually restrict the covariance matrices to be diagonal because the linear combination of diagonal covariance basis Gaussians is capable of modeling the correlations between feature vector elements.

c) Correlation normalizes the contribution of each variable in order to measure only how much the variables are related,rather than also being affected by the scale of the separate variables, while covariance gives some sense of how much two values are linearly related to each other, as well as the scale of these variables.

d) In the absence of a prior knowledge about what form a distribution over the real numbers should take, Gaussian distribution is a good choice for following reasons:

- First, many distributions we wish to model are truly close to being normal distributions. The central limit theorem shows that the sum of many independent random variables is approximately normally distributed. This means that in practice, many complicated systems can be modeled successfully as normally distributed noise, even if the system can be decomposed into parts with more structured behaviour.
- Second, out of all possible probability distributions with the same variance, the Gaussian distribution encodes the maximum amount of uncertainty over the real numbers.

e) Multinoulli distribution is a distribution over a single discrete variable with *k* different states where *k* is finite, whereas multinomial distribution is a distribution over vectors in $\{0, ..., n\}^k$ representing how many times each of the k states is visited when *n* samples are drawn from a multinoulli distribution. Hence, multinoulli distribution is a special case of multinomial distribution where n=1.

f) Multinoulli distribution is parameterized by a vector $p \in [0,1]^k$, where p_i gives the probability of the *i*-th state. These distributions are often used to refer to distributions over categories of objects, so we do not usually assume that state 1 has numerical value 1 and so on. For this reason, we do not usually need to compute the expectation or variance of multinoulli-distributed random variables.

Question 5:

a) Most common form of optimization used in Deep Learning is the gradient descent which belongs to the first order optimization algorithm family.

b) First-order optimization algorithms only use the gradient, whereas second-order optimization algorithms also use the Hessian matrix. Therefore, first order optimization techniques are easy to compute and less time consuming, converging pretty fast on large datasets. In contrast, second-order optimization algorithms are always slower and costly to compute in terms of both time and memory, but they will not get stuck around paths of slow convergence around saddle points whereas former techniques do sometimes get stuck and does not converge.