AI Engineer Training: II
In the Era of Deep Learning

IT21 Learning
Alvin Jin

Weekly AI News

- Samsung opens AI Centre in Toronto
- Trudeau secures Canada's foothold in AI research at MIT
- U.S. to put limits on visas for Chinese student in sensitive technologies, e.g. AI, robotics
- Baidu spins out its Ad business to focus on AI
- Google renames research division as Google AI

Agenda

- Machine Learning Principles
- Case Study: Sentimental Analysis

Machine Learning Categories

- Supervised Learning
  - Learn to map input data to known targets
- Unsupervised Learning
  - Find interesting transformations of input data without any targets
- Reinforcement Learning
  - Agent receives information about its environment, and learns to choose actions that will maximize rewards.

Classic ML Algorithms

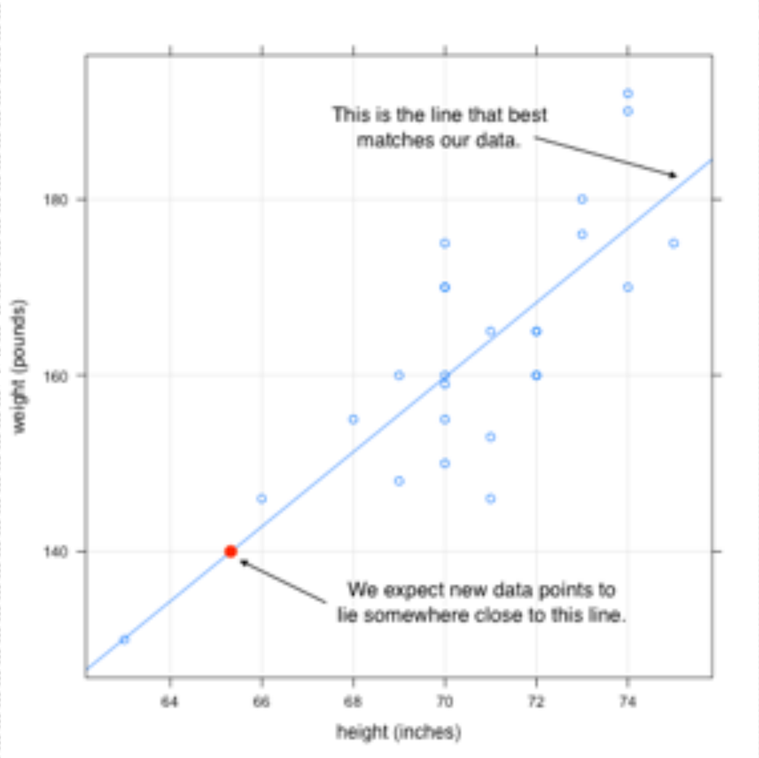
- Classification: functions to predict target classes.
  supervised
- Regression: functions to predict a discrete or continues value.
  supervised
- Clustering: use a distance measure iteratively moving similar items more closely together.
  unsupervised

Machine Learning Model

- The function generated when train ML algorithms on the training dataset.
- Find values of a and b, so f(x)=ax+b matches data points closely.

coefficient

bias



Parameters

- Model Parameters = Weights = Kernel
  - The learnable part of the model that is learned from training data.
  - The values of these parameters before learning starts are initialized randomly
  - Then adjusted towards values that have optimal output.

- Model Hyper-parameters: 不存在在函数中
  - Variables manually set before actually optimizing the model parameters.
  - e.g. Learning Rate, Batch Size, Epoch, etc.

loss function: predict的值和实际值的值。sum(observed - estimated)^2

Case Study: Sentimental Analysis

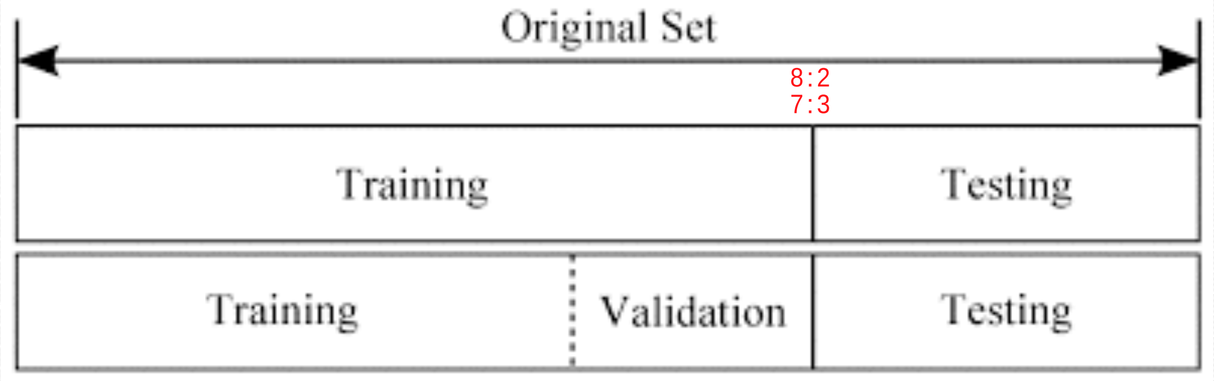
- Binary Classification
  - Each input sample is categorized into two exclusive categories, e.g. positive or negative, sick or not, etc.

- IMDB data set:
  - 50,000 reviews.
  - positive or negative?

Machine Learning Data Sets

- Training data: train new model to learn weights

- Test data: test final model on it for performance estimate, when model is ready.



Validation Data Sets

- Validation set is used to evaluate the performance of the model given the hyper-parameters.

- Based on model's performance to tune a hyper-parameter

Data Processing

- Vectorization
  - Convert raw data into tensors
- Normalization
  - all feature values are in the same range with standard deviation of 1 and mean of 0.
- Feature Engineering
  - make the algorithm work better by applying human knowledge to the data before modelling

Vectorization

- Convert raw data as numeric tensors to feed into neural network
- Matrix operations are magnitudes faster than standard loops, avoid for-loops!
- Approaches:
  - One-hot-encoding
  - Embedding

Model Evaluations

- The goal is to achieve models that generally perform well on never-before-seen data
- Choose appropriate metrics to judge the performance of the models
- Overfitting is the central obstacle

Optimizer

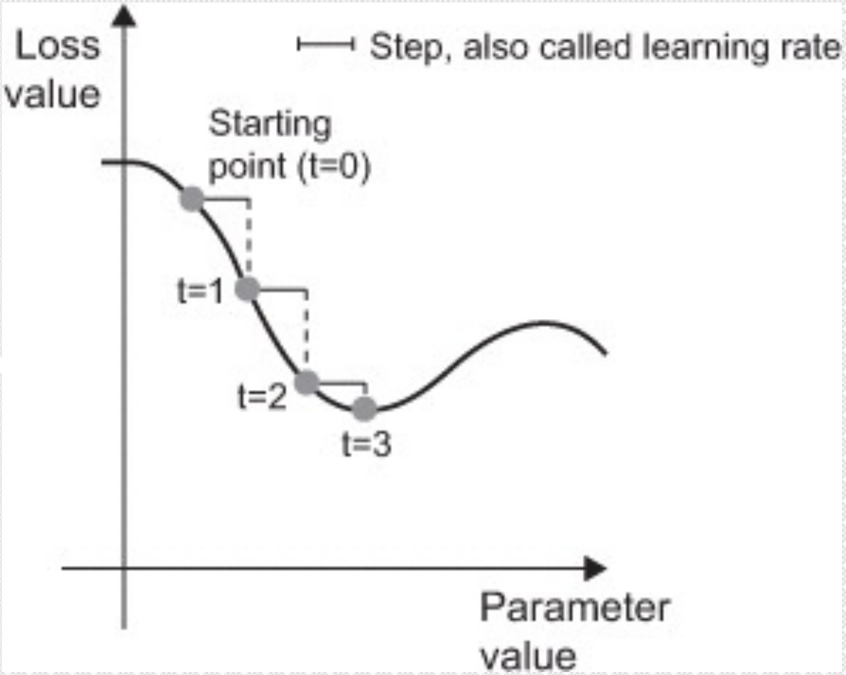
- Gradients of complex functions either vanish or explode as the energy is propagated through the function.

- Popular Optimizers:
  - SGD (Stochastic Gradient Descent)
  - RMSProp (Root Mean Square Propagation) Hinton
  - Momentum
  - Adam (Adaptive Moment Estimation)

Learning Rate

- Controls how much we are adjusting the weights of our network with respect the loss gradient.
- newWeight = oldWeight - learningRate * gradient

0.001, suggested by Hinton



RMSProp

$$E[g^2]_t = 0.9E[g^2]_{t-1} + 0.1g_t^2$$
$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t$$

- Divide the overall learning rate by the square root of the sum of squares of the previous update gradients for a given parameter

- Decreases the step for large gradient to avoid exploding,

- Increases the step for small gradient to avoid vanishing.

Loss Function

- Objective, Loss, Cost, Error function is synonymous.

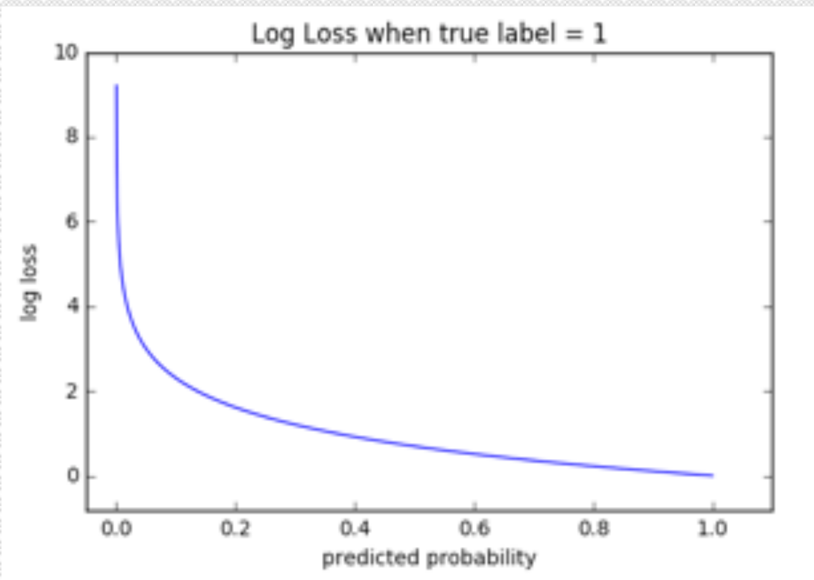
- The function that will get minimized by the Optimizer to optimize your model.

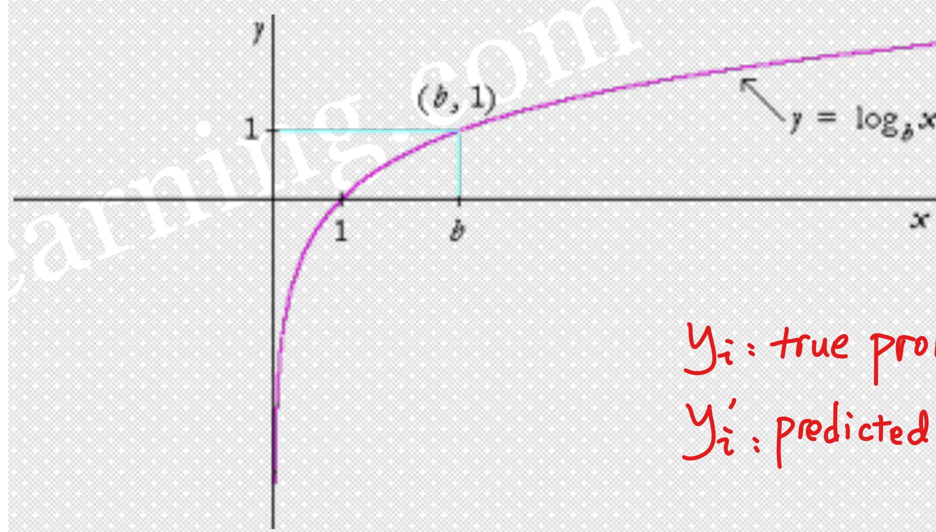
- Widely Used Loss Functions:
  - Cross Entropy (Log Loss)
  - Mean Squared Error (MSE)
  - Mean Absolute Error (MAE)

Binary Cross-Entropy Function

- Output is a probability value between 0 and 1.
- Loss increases as the predicted probability diverges from the actual label.

training loss & training accuracy
validation loss & validation accuracy





y_i : true prob
y_i' : predicted prob

$$H_{y'}(y) := -\sum_i \left(y_i' \log(y_i) + (1 - y_i') \log(1 - y_i)\right)$$

Metrics

- For researchers to judge the performance of models on the validation set after each epoch.

- Classification Metrics
  - Accuracy
  - Logarithmic loss

- Regression Metrics:
  - Mean Absolute Error.
  - Mean Squared Error

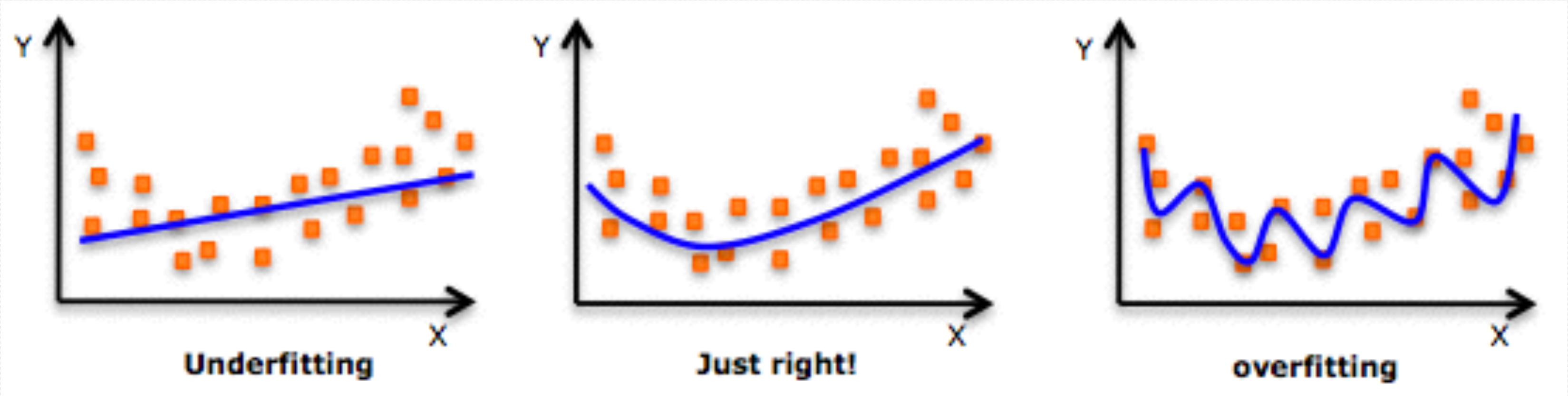
Optimization vs. Generalization

- Optimization
  - The process of adjusting a model to get the best performance on the training data
- Generalization
  - How well the trained model performs on data it has never seen before.

Underfitting vs Overfitting

- Underfitting:
  - The network hasn't yet modelled all relevant patterns in the training data.
  - performs poorly on the training data
- Overfitting:
  - Begin to learn patterns that are specific to the training data, including noise and details
  - performs well on the training data, but not well on the evaluation data

Underfitting vs Overfitting



- Underfitting:
  - Easy to detect given a good performance metric
  - Add new domain-specific features
- Overfitting:
  - Select fewer features
  - Increase the amount of training samples.

Q & A

Normalization

- Take small values
  - Most values are in 0-1 range.
- Homogenous
  - All features should take values in roughly the same range
- Each feature has a standard deviation of 1 and a mean of 0

Standard Deviation

- A measure that is used to quantify the amount of variation of a set of values.
- A low standard deviation: the data points tend to be close to the mean
- A high standard deviation: the data points are spread out over a wider range of values.

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$n =$ The number of data points

$\bar{x} =$ The mean of the x_i

$x_i =$ Each of the values of the data

Feature Engineering

- The process of using domain knowledge of the data by expressing it in a simpler way.
- It is hard, time-consuming, arts rather than science.
- Neural networks are capable of automatically extracting useful features from raw data.
- However, good feature engineering
  - Solve problems elegantly using fewer resources.
  - Solve a problem with far less data.