Machine Learning Basics[1]

STA442/2101 Fall 2018

[1] See last slide for copyright information.

Source

Chapter 5 in *Deep Learning* by Goodfellow, Bengio and Courville

I have copy-pasted so many quotes from this text that this slide show fits the definition of plagiarism.

Machine learning is a form of applied statistics

Machine learning is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions.

Learning

> *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

The term learning can also mean model fitting or parameter estimation.

Tasks

Machine learning tasks are usually described in terms of how the machine learning system should process an example. An ***example*** *is a collection of* ***features*** *that have been quantitatively measured from some object or event that we want the machine learning system to process.*

- Example = datausually what we would call a case
- Feature = variable

Examples of common tasks T

- Classification: In this type of task, the computer program is asked to specify which of k categories some input belongs to. Can be withor without missing inputs (medical diagnosis).
- Regression: In this type of task, the computer program is asked to predict a numerical value given some input.
- Transcription: In this type of task, the machine learning system is asked to observe a relatively unstructured representation of some kind of data and transcribe it into discrete, textual form. For example, in optical character recognition, ...
- Machine translation: In a machine translation task, the input already consists of a sequence of symbols in some language, and the computer program must convert this into a sequence of symbols in another language.

More Examples of tasks

- Anomaly detection: In this type of task, the computer program sifts through a set of events or objects, and flags some of them as being unusual or atypical. An example of an anomaly detection task is credit card fraud detection. *To me, detection of credit card fraud is a classification problem. Another example of anomoly detection is outlier detection.*
- Imputation of missing values.
- Denoising: The machine learning algorithm is given as input a corrupted example obtained by an unknown corruption process from a clean example. The learner must predict the clean example from its corrupted version. *Sounds like measurement error modeling.*
- Density estimation or probability mass function estimation.

Performance measure P

- Performance measures are usually specific to the task. Accuracy is an example.
- "We often refer to the error rate as the expected 0-1 loss. The 0-1 loss on a particular example is 0 if it is correctly classified and 1 if it is not."

  Usually we are interested in how well the machine learning algorithm performs on data that it has not seen before, since this determines how well it will work when deployed in the real world. We therefore evaluate these performance measures using a ***test*** *set of data that is separate from the data used for training the machine learning system.*

Experience E

Machine learning algorithms can be broadly categorized as unsupervised or supervised by what kind of experience they are allowed to have during the learning process. Most of the learning algorithms in this book can be understood as being allowed to experience an entire dataset.

Experience = to process data?

Supervised versus unsupervised learning

Roughly speaking, unsupervised learning involves observing several examples of a random vector x, and attempting to implicitly or explicitly learn the probability distribution $p(x)$, or some interesting properties of that distribution, while supervised learning involves observing several examples of a random vector x and an associated value or vector y, and learning to predict y from x, usually by estimating $p(y|x)$.

- Supervised learning: There is a response variable. (Called a "label" or "target.")
- Unsupervised learning: No response variable.
  - Cluster analysis
  - Principal components
  - Density estimation

Design matrix

A design matrix is a matrix containing a different example in each row. Each column of the matrix corresponds to a different feature. For instance, the Iris dataset contains 150 examples with four features for each example.

- Example = case (there are n cases).
- Feature = variable.

Regression example

"A simple machine learning algorithm: linear regression."

- "The goal is to build a system that can take a vector $\mathbf{x} \in \mathbb{R}^n$ as input and predict the value of a scalar $y \in \mathbb{R}$ as its output."
- "We define the output to be $\widehat{y} = \mathbf{w}^\top \mathbf{x}$, where $\mathbf{w} \in \mathbb{R}^n$ is a vector of **parameters**." *ouch.*
- Test set will be used only for evaluation. *Good.*
- "One way of measuring the performance of the model is to compute the mean squared error of the model on the test set." MSE_{test}
- "Minimize the mean squared error on the training set, MSE_{train} ."
- Then they present the normal equations and say that evaluating $\mathbf{w} = \left(\mathbf{X}^{(\text{train})\top}\mathbf{X}^{(\text{train})}\right)^{-1}\mathbf{X}^{(\text{train})\top}\mathbf{y}^{(\text{train})}$ "constitutes a simple learning algorithm." So "learning" is definitely estimation, or at least curve fitting.

What is a "model?"

- I know what a model is in statistics. It's a set of assertions that implies a probability distribution for the observable data.
- In machine learning, the meaning is slippery – close but not quite the same.
- The "system that can take a vector $\mathbf{x} \in \mathbb{R}^n$ as input and predict the value of a scalar $y \in \mathbb{R}$ as its output" would probably be called a model.
- In statistics, this would be a combination of a model and an estimator.

Generalization

- Generalization: The ability to perform well on previously unobserved inputs.
- With access to a training set, we can compute some error measure on the training set called the **training error**, and we reduce this training error. *This is just optimization.*
- The generalization error is defined as the expected value of the error on a new input. *Good, that's clear.*
- We want the generalization error, also called the **test error**, to be low as well. *That is, we want generalization error as well as training error to be low.*
- There's a theorem that generalization error is greater than or equal to (expected) training error

Objectives

- Make the training error small.
- Make the gap between training and test error small.

Data generating distribution
The i.i.d. assumptions

To get anywhere, even the machine learning people have to make some assumptions.

- Assume training set and test set are independent.
- Examples are independent within sets.
- Both training set and test set come from a common **data generating distribution distribution** denoted p_{data}.

Over and underfittng

- Underfitting occurs when the model is not able to obtain a sufficiently low error value on the training set.
- Overfitting occurs when the gap between the training error and test error is too large.
- *Notice how empirical this is compared to the statistical formulation.*

Model capacity

- A model's **capacity** is its ability to fit a wide variety of functions.
- "Models with low capacity may struggle to fit the training set."
- "Models with high capacity can overfit by memorizing properties of the training set that do not serve them well on the test set."
- **Hypothesis space**: The set of functions that the learning algorithm is allowed to select as being the solution. *Polynomial regression example.*
- "The model specifies which family of functions the learning algorithm can choose from when varying the parameters in order to reduce a training objective.This is called the **representational capacity** of the model."
- *So the representational capacity appears to be determined by the hypothesis space.*

There are theoretical results

> *The most important results in statistical learning theory show that the discrepancy between training error and generalization error is bounded from above by a quantity that grows as the model capacity grows but shrinks as the number of training examples increases.*

The authors note that these bounds are very loose and seldom used in practice.

k nearest neighbor regression

- Capacity of the model grows with the size of the data set.
- "This algorithm is able to achieve the minimum possible training error on any regression dataset."

Bayes error

- The ideal model is an oracle that simply knows the true probability distribution that generates the data.
- The error incurred by an oracle making predictions from the true distribution $p(\mathbf{x}, y)$ is called the Bayes error.
- *I have no idea why.*

The No Free Lunch Theorem

> *The no free lunch theorem for machine learning (Wolpert, 1996) states that, averaged over all possible data generating distributions, every classification algorithm has the same error rate when classifying previously unobserved points.*

So you need to make some assumptions about the probability distributions that might reasonably be encountered in practice.

Regularization

- The no free lunch theorem implies that we must design our machine learning algorithms to perform well on a specific task.
- Choose functions in the hypothesis space well.
- We can also give a learning algorithm a preference for one solution in its hypothesis space over another.
- This is called **regularization**.
- Weight decay example in regression: Minimize

$$J(\mathbf{w}) = MSE_{\text{train}} + \lambda \mathbf{w}^\top \mathbf{w}$$

- *I don't know about the crazy vocabulary, but there is some freedom here that is harder to find in standard statistical practice.*

Hyperparameters

- Most machine learning algorithms have several settings that we can use to control the behavior of the learning algorithm.
- These settings are called hyperparameters.
- The values of hyperparameters are not adapted by the learning algorithm itself.
- *This is very different from the meaning of hyperparameter as I understand it.*

Validation set
A useful concept

- *Want to search around in the hypothesis space to locate the best model, but that could result in over-fitting.*
- *Can't peek at the test data.*
- Split the training data, *yes the training data*, into two disjoint subsets.
- One of these subsets is used to learn the parameters.
- The other subset is our validation set, used to estimate the generalization error during or after training, allowing for the hyperparameters to be updated accordingly.
- The subset of data used to learn the parameters is still typically called the training set, even though this may be confused with the larger pool of data used for the entire training process.

k-fold cross validation

- You don't have enough data to split into a training set and a test set.
- Split into k disjoint subsets. Try to predict each one in turn using the rest of the data as a training sample.
- Average the results.
- The variance of such an average is hard to estimate well.

Standard statistical ideas

Sometimes with a strange twist

- Sometimes there really are unknown parameters and they do maximum likelihood.
- They still refer to estimation as prediction, most of the time.
- "In machine learning experiments, it is common to say that algorithm A is better than algorithm B if the upper bound of the 95% confidence interval for the error of algorithm A is less than the lower bound of the 95% confidence interval for the error of algorithm B."
- *Why not just test?*

Bayesian Statistics

- Pretty standard treatment.
- Predictive density: Integrate out the parameter using the posterior distribution.
- Regression example with a conjugate prior: "The Bayesian estimate provides a covariance matrix, showing how likely all the different values of $\mathbf{w}$ are, rather than providing only a the estimate ..."
- "Maximum A Posteriori (MAP) Estimation" means estimate using the posterior mode.

Support vector machines

- Method for binary classification.
- It looks pretty strong. This is new to me, I think.
- Predict Yes for test data $\mathbf{x}$ when $\mathbf{w}^\top \mathbf{x} + b$ is positive.
- Predict No when $\mathbf{w}^\top \mathbf{x} + b$ is negative.
- Replace $\mathbf{w}^\top \mathbf{x}$ with $b + \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}) \cdot \phi(\mathbf{x}^{(i)})$,
- Where $\mathbf{x}^{(i)}$ is a vector of training data and the dot product is very general.

Stochastic Gradient Descent

- A recurring problem in machine learning is that large training sets are necessary for good generalization, but large training sets are also more computationally expensive.
- The cost function used by a machine learning algorithm often decomposes as a sum over training examples of some per-example loss function. *The minus log likelihood is a sum over observations.*
- The sample size can be huge.
- Calculating the big sum (of derivatives) can take a lot of computation.
- So just compute the gradient on a random sample. Random = stochastic.
- Go downhill, a little randomly.
- It's no big deal, but why not just do the whole task on a random sample of the data?

Recipe for a machine learning algorithm

Combine

- Dataset
- Cost function
- Optimization procedure
- Model

Copyright Information