

# CSE 6740 A/ISyE 6740: Computational Data Analysis: Introductory lecture

Nisha Chandramoorthy

August 22, 2023

# Course logistics

- ▶ Instructor: Nisha Chandramoorthy.

# Course logistics

- ▶ Instructor: Nisha Chandramoorthy.
- ▶ Interested in teaching and learning about foundations of machine learning

# Course logistics

- ▶ Instructor: Nisha Chandramoorthy.
- ▶ Interested in teaching and learning about foundations of machine learning
- ▶ 6 TAs: Akpevwe Ojameruaye, Atharva Ketkar, Chengrui Li, Darryl Jacob,, Mithilesh Vaidya, and Yusen Su

# Course logistics

- ▶ Instructor: Nisha Chandramoorthy.
- ▶ Interested in teaching and learning about foundations of machine learning
- ▶ 6 TAs: Akpevwe Ojameruaye, Atharva Ketkar, Chengrui Li, Darryl Jacob,, Mithilesh Vaidya, and Yusen Su
- ▶ Lectures: TR 12:30-1:45 pm. OH: 30 minutes after. Location: East Architecture 123.

# Course logistics

- ▶ Instructor: Nisha Chandramoorthy.
- ▶ Interested in teaching and learning about foundations of machine learning
- ▶ 6 TAs: Akpevwe Ojameruaye, Atharva Ketkar, Chengrui Li, Darryl Jacob,, Mithilesh Vaidya, and Yusen Su
- ▶ Lectures: TR 12:30-1:45 pm. OH: 30 minutes after. Location: East Architecture 123.
- ▶ Grade: 4 homeworks (30%), 2 midterms (30%), final project (40%)

# Course logistics

- ▶ Instructor: Nisha Chandramoorthy.
- ▶ Interested in teaching and learning about foundations of machine learning
- ▶ 6 TAs: Akpevwe Ojameruaye, Atharva Ketkar, Chengrui Li, Darryl Jacob,, Mithilesh Vaidya, and Yusen Su
- ▶ Lectures: TR 12:30-1:45 pm. OH: 30 minutes after. Location: East Architecture 123.
- ▶ Grade: 4 homeworks (30%), 2 midterms (30%), final project (40%)
- ▶ **Canvas** (see syllabus), **gradescope**, **Piazza**, **Github**

# Machine learning and data mining: what are they?

- ▶ “Automated detection of meaningful patterns in data” - Shalev-Shwartz and Ben-David.



# Machine learning and data mining: what are they?

- ▶ “Automated detection of meaningful patterns in data” - Shalev-Shwartz and Ben-David.
- ▶ Goal in this class: understand the foundations (“why”s and “how”s) of ML

# Machine learning and data mining: what are they?

- ▶ “Automated detection of meaningful patterns in data” - Shalev-Shwartz and Ben-David.
- ▶ Goal in this class: understand the foundations (“why”s and “how”s) of ML
- ▶ ML = Compute + data

# Machine learning and data mining: what are they?

- ▶ “Automated detection of meaningful patterns in data” - Shalev-Shwartz and Ben-David.
- ▶ Goal in this class: understand the foundations (“why”s and “how”s) of ML
- ▶ ML = Compute + data
- ▶ Compute: Optimization, representation/models

# Machine learning and data mining: what are they?

- ▶ “Automated detection of meaningful patterns in data” - Shalev-Shwartz and Ben-David.
- ▶ Goal in this class: understand the foundations (“why”s and “how”s) of ML
- ▶ ML = Compute + data
- ▶ Compute: Optimization, representation/models
- ▶ Data: distributions, features/compression, statistics

# Categorizations of learning

- ▶ Supervised, unsupervised, self-supervised, semi-supervised

# Categorizations of learning

- ▶ Supervised, unsupervised, self-supervised, semi-supervised
- ▶ Supervised: using *experience* (training data) to learn

# Categorizations of learning

- ▶ Supervised, unsupervised, self-supervised, semi-supervised
- ▶ Supervised: using *experience* (training data) to learn
- ▶ Unsupervised: using *data* to identify patterns, match distributions?

# Categorizations of learning

- ▶ Supervised, unsupervised, self-supervised, semi-supervised
- ▶ Supervised: using *experience* (training data) to learn
- ▶ Unsupervised: using *data* to identify patterns, match distributions?
- ▶ Mode of learning and testing are different



# Categorizations of learning

- ▶ Supervised, unsupervised, self-supervised, semi-supervised
- ▶ Supervised: using *experience* (training data) to learn
- ▶ Unsupervised: using *data* to identify patterns, match distributions?
- ▶ Mode of learning and testing are different
- ▶ You sample peaches across several grocery stores in Atlanta. Now you are given a new peach of unknown origins. Can you tell if it would taste good?

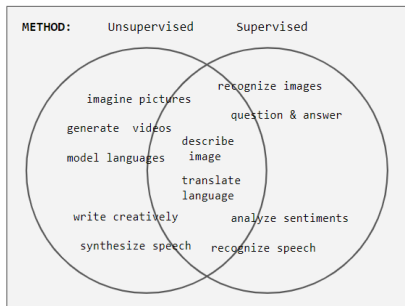
# Supervised, unsupervised, in-between, semi-supervised, self-supervised...



Cat



Dog



- Many modern tasks require both modes of learning

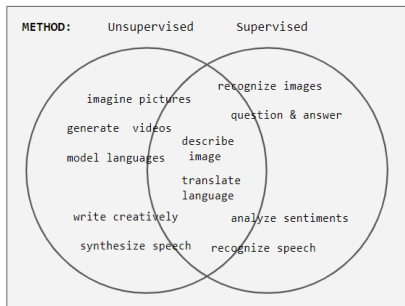
# Supervised, unsupervised, in-between, semi-supervised, self-supervised...



Cat



Dog



- ▶ Many modern tasks require both modes of learning
- ▶  $5 * 9 = -4$ ,  $4 + 10 = 6$ ,  $8 * 7 = 1$ ,  $5 + 2 = -3$ ,  $(3 * 5) + 6 = ?$

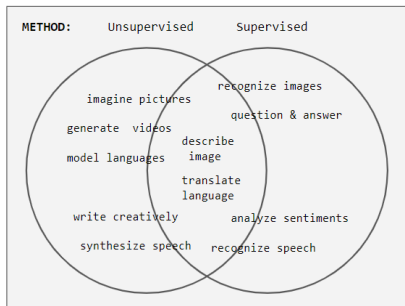
# Supervised, unsupervised, in-between, semi-supervised, self-supervised...



Cat



Dog



- ▶ Many modern tasks require both modes of learning
- ▶  $5 * 9 = -4$ ,  $4 + 10 = 6$ ,  $8 * 7 = 1$ ,  $5 + 2 = -3$ ,  $(3 * 5) + 6 = ?$
- ▶ You are feeling sleepy this afternoon. You order a ...

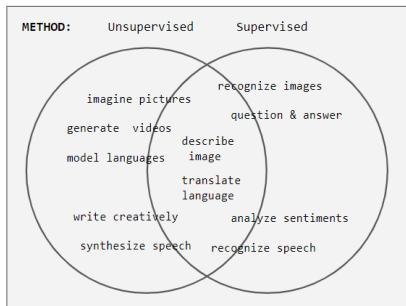
# Supervised, unsupervised, in-between, semi-supervised, self-supervised...



Cat



Dog



- ▶ Many modern tasks require both modes of learning
- ▶  $5 * 9 = -4$ ,  $4 + 10 = 6$ ,  $8 * 7 = 1$ ,  $5 + 2 = -3$ ,  $(3 * 5) + 6 = ?$
- ▶ You are feeling sleepy this afternoon. You order a ...
- ▶ New research frontier for theory: understand how and why large ML models work the way they do?

# Models/representations, algorithms, statistical principles

- ▶ how to make and test conjectures about how large language models (LLMs) learn?

# Models/representations, algorithms, statistical principles

- ▶ how to make and test conjectures about how large language models (LLMs) learn?
- ▶ “how to train them better (more efficiently)” – number of practical questions perhaps benefit from theory

# Models/representations, algorithms, statistical principles

- ▶ how to make and test conjectures about how large language models (LLMs) learn?
- ▶ “how to train them better (more efficiently)” – number of practical questions perhaps benefit from theory
- ▶ AI safety, fair and ethical use, combining with other domain knowledge (e.g., physics, chemistry etc).... and many more!



# Models/representations, algorithms, statistical principles

- ▶ how to make and test conjectures about how large language models (LLMs) learn?
- ▶ “how to train them better (more efficiently)” – number of practical questions perhaps benefit from theory
- ▶ AI safety, fair and ethical ethical use, combining with other domain knowledge (e.g., physics, chemistry etc).... and many more!
- ▶ Perhaps biggest contribution advance to LLMs: transformers and their training.

# (partial) History - trace back from transformers

(source:Wikipedia)

- ▶ Transformer architecture: 2017, Google Brain [Vaswani et al]
- ▶ Deep learning, unsupervised learning 2010s (e.g., GANs 2014)...
- ▶ ImageNet: 2009, Fei Fei Li
- ▶ Long-short term memory (LSTM) architecture: 1997, [Hochreiter and Schmidhuber]
- ▶ Convolutional NNs: (inspired from) 1979 work by [Fukushima]; Recurrent neural networks: 1982 [Hopfield]
- ▶ ...
- ▶ Automatic Differentiation: 1970 [Linnainmaa]
- ▶ ...
- ▶ First neural networks: 1950s [Minsky and others]

# Supervised learning framework

- ▶ Distribution  $\mathcal{D}$  over the joint distribution of random variables  $Z = (X, Y)$ , where  $X$  is an input and  $Y$  is a label/output.

# Supervised learning framework

- ▶ Distribution  $\mathcal{D}$  over the joint distribution of random variables  $Z = (X, Y)$ , where  $X$  is an input and  $Y$  is a label/output.
- ▶ Labeled training data,  $S = \{z_i = (x_i, y_i)\}, 1 \leq i \leq m$ . Generally iid from  $\mathcal{D}^m$ .

# Supervised learning framework

- ▶ Distribution  $\mathcal{D}$  over the joint distribution of random variables  $Z = (X, Y)$ , where  $X$  is an input and  $Y$  is a label/output.
- ▶ Labeled training data,  $S = \{z_i = (x_i, y_i)\}, 1 \leq i \leq m$ . Generally iid from  $\mathcal{D}^m$ .
- ▶ Learner's output: predicted function or hypothesis, a transformation  $h$  from  $X$  to  $Y$ .

# Supervised learning framework

- ▶ Distribution  $\mathcal{D}$  over the joint distribution of random variables  $Z = (X, Y)$ , where  $X$  is an input and  $Y$  is a label/output.
- ▶ Labeled training data,  $S = \{z_i = (x_i, y_i)\}, 1 \leq i \leq m$ . Generally iid from  $\mathcal{D}^m$ .
- ▶ Learner's output: predicted function or hypothesis, a transformation  $h$  from  $X$  to  $Y$ .
- ▶ Loss function, measure of risk:  $\ell(z, h) \in \mathbb{R}$ . e.g.,  $\ell(z) = \mathbb{1}_{h(x) \neq y}$ . (classification)

# Supervised learning framework

- ▶ Distribution  $\mathcal{D}$  over the joint distribution of random variables  $Z = (X, Y)$ , where  $X$  is an input and  $Y$  is a label/output.
- ▶ Labeled training data,  $S = \{z_i = (x_i, y_i)\}, 1 \leq i \leq m$ . Generally iid from  $\mathcal{D}^m$ .
- ▶ Learner's output: predicted function or hypothesis, a transformation  $h$  from  $X$  to  $Y$ .
- ▶ Loss function, measure of risk:  $\ell(z, h) \in \mathbb{R}$ . e.g.,  $\ell(z) = \mathbb{1}_{h(x) \neq y}$ . (classification)
- ▶ Generalization error or risk:

$$R(h) = E_{z \sim \mathcal{D}} \ell(z)$$

# Empirical Risk Minimization

- ▶ Take classifier  $h$ ,  $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$ .



# Empirical Risk Minimization

- ▶ Take classifier  $h$ ,  $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$ .
- ▶ When you have finite amount of data

$$\hat{R}_S(h) = \frac{1}{m} \sum_{z \in S} \ell(z) = \frac{|\{z \in S : h(x) \neq y\}|}{m}.$$

# Empirical Risk Minimization

- ▶ Take classifier  $h$ ,  $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$ .
- ▶ When you have finite amount of data

$$\hat{R}_S(h) = \frac{1}{m} \sum_{z \in S} \ell(z) = \frac{|\{z \in S : h(x) \neq y\}|}{m}.$$

- ▶ ERM: find  $h$  that minimizes  $\hat{R}_S(h)$ .

# Empirical Risk Minimization

- ▶ Take classifier  $h$ ,  $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$ .
- ▶ When you have finite amount of data

$$\hat{R}_S(h) = \frac{1}{m} \sum_{z \in S} \ell(z) = \frac{|\{z \in S : h(x) \neq y\}|}{m}.$$

- ▶ ERM: find  $h$  that minimizes  $\hat{R}_S(h)$ .
- ▶ Theory of supervised learning suggests that ERM leads to small generalization error with high probability

# Empirical Risk Minimization

- ▶ Take classifier  $h$ ,  $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$ .
- ▶ When you have finite amount of data

$$\hat{R}_S(h) = \frac{1}{m} \sum_{z \in S} \ell(z) = \frac{|\{z \in S : h(x) \neq y\}|}{m}.$$

- ▶ ERM: find  $h$  that minimizes  $\hat{R}_S(h)$ .
- ▶ Theory of supervised learning suggests that ERM leads to small generalization error with high probability
- ▶ Now we will prove this result formally.

# Empirical Risk Minimization

- ▶ Take classifier  $h$ ,  $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$ .
- ▶ When you have finite amount of data

$$\hat{R}_S(h) = \frac{1}{m} \sum_{z \in S} \ell(z) = \frac{|\{z \in S : h(x) \neq y\}|}{m}.$$

- ▶ ERM: find  $h$  that minimizes  $\hat{R}_S(h)$ .
- ▶ Theory of supervised learning suggests that ERM leads to small generalization error with high probability
- ▶ Now we will prove this result formally.
- ▶ Next time: Linear models.

# Finite hypothesis classes

- ▶ ERM rules can lead to bad generalization

# Finite hypothesis classes

- ▶ ERM rules can lead to bad generalization
- ▶ Why? Overfitting

# Finite hypothesis classes

- ▶ ERM rules can lead to bad generalization
- ▶ Why? Overfitting
- ▶ Example



# Finite hypothesis classes

- ▶ ERM rules can lead to bad generalization
- ▶ Why? Overfitting
- ▶ Example
- ▶ One way to reduce overfitting: inductive bias. Choose  $\mathcal{H}$  based on prior knowledge.

# Finite hypothesis classes

- ▶ ERM rules can lead to bad generalization
- ▶ Why? Overfitting
- ▶ Example
- ▶ One way to reduce overfitting: inductive bias. Choose  $\mathcal{H}$  based on prior knowledge.
- ▶ Later: when does memorization of training data lead to good generalization?

# Finite hypothesis classes

- ▶ ERM rules can lead to bad generalization
- ▶ Why? Overfitting
- ▶ Example
- ▶ One way to reduce overfitting: inductive bias. Choose  $\mathcal{H}$  based on prior knowledge.
- ▶ Later: when does memorization of training data lead to good generalization?
- ▶ Now: simple case of finite  $\mathcal{H}$ .