Course Goals

Be introduced to some of the core problems and solutions of ML (big picture)
Machine Learning Framework: Learning

- instance 1
- instance 2
- instance 3
- instance 4

Machine Learning Predictor

Evaluator

Gold/correct labels

score

give feedback to the predictor

Extra-knowledge

instances are typically examined independently

Gold/correct labels give feedback to the predictor
Course Goals

Be introduced to some of the core problems and solutions of ML (big picture)

Learn different ways that success and progress can be measured in ML
Machine Learning Framework: Learning

Instance 1
Instance 2
Instance 3
Instance 4

Machine Learning Predictor

Gold/correct labels

Evaluator

score

sometimes: the function we can learn != the function we want to learn

give feedback to the predictor

instances are typically examined independently

Extra-knowledge
Optimize Empirical Risk of (Surrogate) Loss

$$\arg\min_h \sum_{i=1}^{N} \ell(y_i, h_\theta(x_i))$$

$$\nabla_\theta F = \sum_i \frac{\partial \ell(y_i, \hat{y} = h_\theta(x_i))}{\partial \hat{y}} \nabla_\theta h_\theta(x_i)$$

empirical risk minimization

approximate loss in a computable way
Course Goals

Be introduced to some of the core problems and solutions of ML (big picture)

Learn different ways that success and progress can be measured in ML

Relate to statistics, AI [671], and specialized areas (e.g., NLP [673] and CV [691])

Implement ML programs
Course Goals

Be introduced to some of the core problems and solutions of ML (big picture)
Learn different ways that success and progress can be measured in ML
Relate to statistics, AI [671], and specialized areas (e.g., NLP [673] and CV [691])
Implement ML programs
Read and analyze research papers
Practice your (written) communication skills
A Terminology Buffet

Classification
Regression
Clustering

Fully-supervised
Semi-supervised
Un-supervised

the task: what kind of problem are you solving?
the data: amount of human input/number of labeled examples
the approach: how any data are being used

Probabilistic
Generative
Conditional
Spectral

Neural
Memory-based
Exemplar
...
A Terminology Buffet

- **Classification**
  - Regression
  - Clustering

- **Fully-supervised**

- **Semi-supervised**

- **Un-supervised**

**the task**: what kind of problem are you solving?

**the data**: amount of human input/number of labeled examples

**the approach**: how any data are being used
Course Overview (Part 1)

Basics of Probability
- Requirements to be a distribution ("proportional to", $\propto$)
- Definitions of conditional probability, joint probability, and independence
- Bayes rule, (probability) chain rule
- Expectation (of a random variable & function)

Empirical Risk Minimization
- Gradient Descent
- Loss Functions: what is it, what does it measure, and what are some computational difficulties with them?
- Regularization: what is it, how does it work, and why might you want it?

Tasks (High Level)
- Data set splits: training vs. dev vs. test
- Classification: Posterior decoding/MAP classifier
- Classification evaluations: accuracy, precision, recall, and F scores
- Regression (vs. classification)
- Comparing supervised vs. Unsupervised Learning and their tradeoffs: why might you want to use one vs. the other, and what are some potential issues?
- Clustering: high-level goal/task, K-means as an example
- Tradeoffs among clustering evaluations

Linear Models
- Basic form of a linear model (classification or regression)
- Perceptron (simple vs. other variants, like averaged or voted)
- When you should use perceptron (what are its assumptions?)
- Perceptron as SGD

Maximum Entropy Models
- Meanings of feature functions and weights
- How to learn the weights: gradient descent
- Meaning of the maxent gradient

Neural Networks
- Relation to linear models and maxent
- Types (feedforward, CNN, RNN)
- Learning representations (e.g., "feature maps")
- What is a convolution (e.g., 1D vs 2D, high-level notions of why you might want to change padding or the width)
- How to learn: gradient descent, backprop
- Common activation functions
- Neural network regularization

Dimensionality Reduction
- What is the basic task & goal in dimensionality reduction?
- Dimensionality reduction tradeoffs: why might you want to, and what are some potential issues?
- Linear Discriminant Analysis vs. Principal Component Analysis: what are they trying to do, how are they similar, how do they differ?

Kernel Methods & SVMs
- Feature expansion and kernels
- Two views: maximizing a separating hyperplane margin vs. loss optimization (norm minimization)
- Non-separability & slack
- Sub-gradients
A Terminology Buffet

- **Classification**
- **Regression**
- **Clustering**

**Fully-supervised**

**Semi-supervised**

**Un-supervised**

**Probabilistic**

**Generative**

**Memory-based**

**Exemplar**

**Spectral**

**Neural**

**Conditional**

**…**

**the task**: what kind of problem are you solving?

**the data**: amount of human input/number of labeled examples

**the approach**: how any data are being used

*what we’ve currently sampled...*  *what we’ll be sampling next...*
Course Overview (Part 2)

Basics of Probability
- Requirements to be a distribution ("proportional to", $\propto$)
- Definitions of conditional probability, joint probability, and independence
- Bayes rule, (probability) chain rule
- Expectation (of a random variable & function)

Empirical Risk Minimization
- Gradient Descent
- Loss Functions: what is it, what does it measure, and what are some computational difficulties with them?
- Regularization: what is it, how does it work, and why might you want it?

Tasks (High Level)
- Data set splits: training vs. dev vs. test
- Classification: Posterior decoding/MAP classifier
- Classification evaluations: accuracy, precision, recall, and F scores
- Regression (vs. classification)
- Comparing supervised vs. Unsupervised Learning and their tradeoffs: why might you want to use one vs. the other, and what are some potential issues?
- Clustering: high-level goal/task, K-means as an example
- Tradeoffs among clustering evaluations

Linear Models
- Basic form of a linear model (classification or regression)
- Perceptron (simple vs. other variants, like averaged or voted)
- When you should use perceptron (what are its assumptions?)
- Perceptron as SGD

Maximum Entropy Models
- Meanings of feature functions and weights
- How to learn the weights: gradient descent
- Meaning of the maxent gradient

Neural Networks
- Relation to linear models and maxent
- Types (feedforward, CNN, RNN)
- Learning representations (e.g., "feature maps")
- What is a convolution (e.g., 1D vs 2D, high-level notions of why you might want to change padding or the width)
- How to learn: gradient descent, backprop
- Common activation functions
- Neural network regularization

Dimensionality Reduction
- What is the basic task & goal in dimensionality reduction?
- Dimensionality reduction tradeoffs: why might you want to, and what are some potential issues?
- Linear Discriminant Analysis vs. Principal Component Analysis: what are they trying to do, how are they similar, how do they differ?

Kernel Methods & SVMs
- Feature expansion and kernels
- Two views: maximizing a separating hyperplane margin vs. loss optimization (norm minimization)
- Non-separability & slack
- Sub-gradients

leveraging large, unannotated data
Course Overview (Part 2)

Basics of Probability
- Requirements to be a distribution ("proportional to", $\propto$)
- Definitions of conditional probability, joint probability, and independence
- Bayes rule, (probability) chain rule
- Expectation (of a random variable & function)

Empirical Risk Minimization
- Gradient Descent
- Loss Functions: what is it, what does it measure, and what are some computational difficulties with them?
- Regularization: what is it, how does it work, and why might you want it?

Tasks (High Level)
- Data set splits: training vs. dev vs. test
- Classification: Posterior decoding/MAP classifier
- Classification evaluations: accuracy, precision, recall, and F-scores
- Regression (vs. classification)
- Comparing supervised vs. Unsupervised Learning and their tradeoffs: why might you want to use one vs. the other, and what are some potential issues?
- Clustering: high-level goal/task, K-means as an example
- Tradeoffs among clustering evaluations

Linear Models
- Basic form of a linear model (classification or regression)
- Perceptron (simple vs. other variants, like averaged or voted)
- When you should use perceptron (what are its assumptions?)
- Perceptron as SGD

Maximum Entropy Models
- Meanings of feature functions and weights
- How to learn the weights: gradient descent
- Meaning of the maxent gradient

Neural Networks
- Relation to linear models and maxent
- Types (feedforward, CNN, RNN)
- Learning representations (e.g., "feature maps")
- What is a convolution (e.g., 1D vs 2D, high-level notions of why you might want to change padding or the width)
- How to learn: gradient descent, backprop
- Common activation functions
- Neural network regularization

Dimensionality Reduction
- What is the basic task & goal in dimensionality reduction?
- Dimensionality reduction tradeoffs: why might you want to, and what are some potential issues?
- Linear Discriminant Analysis vs. Principal Component Analysis: what are they trying to do, how are they similar, how do they differ?

Kernel Methods & SVMs
- Feature expansion and kernels
- Two views: maximizing a separating hyperplane margin vs. loss optimization (norm minimization)
- Non-separability & slack
- Sub-gradients

leveraging large, unannotated data
Course Overview (Part 2)

Basics of Probability
- Requirements to be a distribution ("proportional to", \(\propto\))
- Definitions of conditional probability, joint probability, and independence
- Bayes rule, (probability) chain rule
- Expectation (of a random variable & function)

Empirical Risk Minimization
- Gradient Descent
- Loss Functions: what is it, what does it measure, and what are some computational difficulties with them?
- Regularization: what is it, how does it work, and why might you want it?

Tasks (High Level)
- Data set splits: training vs. dev vs. test
- Classification: Posterior decoding/MAP classifier
- Classification evaluations: accuracy, precision, recall, and F scores
- Regression (vs. classification)
- Comparing supervised vs. Unsupervised Learning and their tradeoffs: why might you want to use one vs. the other, and what are some potential issues?
- Clustering: high-level goal/task, K-means as an example
- Tradeoffs among clustering evaluations

Linear Models
- Basic form of a linear model (classification or regression)
- Perceptron (simple vs. other variants, like averaged or voted)
- When you should use perceptron (what are its assumptions?)

Perceptron as SGD

Maximum Entropy Models
- Meanings of feature functions and weights
- How to learn the weights: gradient descent
- Meaning of the maxent gradient

Neural Networks
- Relation to linear models and maxent
- Types (feedforward, CNN, RNN)
- Learning representations (e.g., "feature maps")
- What is a convolution (e.g., 1D vs 2D, high-level notions of why you might want to change padding or the width)
- How to learn: gradient descent, backprop
- Common activation functions
- Neural network regularization

Dimensionality Reduction
- What is the basic task & goal in dimensionality reduction?
- Dimensionality reduction tradeoffs: why might you want to, and what are some potential issues?
- Linear Discriminant Analysis vs. Principal Component Analysis: what are they trying to do, how are they similar, how do they differ?

Kernel Methods & SVMs
- Feature expansion and kernels
- Two views: maximizing a separating hyperplane margin vs. loss optimization (norm minimization)
- Non-separability & slack
- Sub-gradents

Latent Variable Models
- Meaning: what’s a latent variable?
- General problem: latent variables are (often) not labeled (or difficult)
- General algorithm: expectation maximization
- Problem EM optimizes (and what it doesn’t)
Course Overview (Part 2)

Basics of Probability
- Requirements to be a distribution ("proportional to", ∝)
- Definitions of conditional probability, joint probability, and independence
- Bayes rule, (probability) chain rule
- Expectation (of a random variable & function)

Empirical Risk Minimization
- Gradient Descent
- Loss Functions: what is it, what does it measure, and what are some computational difficulties with them?
- Regularization: what is it, how does it work, and why might you want it?

Tasks (High Level)
- Data set splits: training vs. dev vs. test
- Classification: Posterior decoding/MAP classifier
- Classification evaluations: accuracy, precision, recall, and F scores
- Regression (vs. classification)
- Comparing supervised vs. Unsupervised Learning and their tradeoffs: why might you want to use one vs. the other, and what are some potential issues?
- Clustering: high-level goal/task, K-means as an example
- Tradeoffs among clustering evaluations

Linear Models
- Basic form of a linear model (classification or regression)
- Perceptron (simple vs. other variants, like averaged or voted)
- When you should use perceptron (what are its assumptions?)

Perceptron as SGD

Maximum Entropy Models
- Meanings of feature functions and weights
- How to learn the weights: gradient descent
- Meaning of the maxent gradient

Neural Networks
- Relation to linear models and maxent
- Types (feedforward, CNN, RNN)
- Learning representations (e.g., "feature maps")
- What is a convolution (e.g., 1D vs 2D, high-level notions of why you might want to change padding or the width)
- How to learn: gradient descent, backprop
- Common activation functions
- Neural network regularization

Dimensionality Reduction
- What is the basic task & goal in dimensionality reduction?
- Dimensionality reduction tradeoffs: why might you want to, and what are some potential issues?
- Linear Discriminant Analysis vs. Principal Component Analysis: what are they trying to do, how are they similar, how do they differ?

Kernel Methods & SVMs
- Feature expansion and kernels
- Two views: maximizing a separating hyperplane margin vs. loss optimization (norm minimization)
- Non-separability & slack
- Sub-gradients

Latent Variable Models
- Meaning: what’s a latent variable?
- General problem: latent variables are (often) not labeled (or difficult)
- General algorithm: expectation maximization
- Problem EM optimizes (and what it doesn’t)

Distributions as Graphs
- Directed graphical models: Bayesian networks
- Undirected graphical models: CRFs, MRFs
- Factor graphs
- Message passing: sum-product
Course Overview (Part 2)

Basics of Probability
- Requirements to be a distribution ("proportional to", ∝)
- Definitions of conditional probability, joint probability, and independence
- Bayes rule, (probability) chain rule
- Expectation (of a random variable & function)

Empirical Risk Minimization
- Gradient Descent
- Loss Functions: what is it, what does it measure, and what are some computational difficulties with them?
- Regularization: what is it, how does it work, and why might you want it?

Tasks (High Level)
- Data set splits: training vs. dev vs. test
- Classification: Posterior decoding/MAP classifier
- Classification evaluations: accuracy, precision, recall, and F scores
- Regression (vs. classification)
- Comparing supervised vs. Unsupervised Learning and their tradeoffs: why might you want to use one vs. the other, and what are some potential issues?
- Clustering: high-level goal/task, K-means as an example
- Tradeoffs among clustering evaluations

Linear Models
- Basic form of a linear model (classification or regression)
- Perceptron (simple vs. other variants, like averaged or voted)
- When you should use perceptron (what are its assumptions?)

Perceptron as SGD

Maximum Entropy Models
- Meanings of feature functions and weights
- How to learn the weights: gradient descent
- Meaning of the maxent gradient

Neural Networks
- Relation to linear models and maxent
- Types (feedforward, CNN, RNN)
- Learning representations (e.g., "feature maps")
- What is a convolution (e.g., 1D vs 2D, high-level notions of why you might want to change padding or the width)
- How to learn: gradient descent, backprop
- Common activation functions
- Neural network regularization

Dimensionality Reduction
- What is the basic task & goal in dimensionality reduction?
- Dimensionality reduction tradeoffs: why might you want to, and what are some potential issues?
- Linear Discriminant Analysis vs. Principal Component Analysis: what are they trying to do, how are they similar, how do they differ?

Kernel Methods & SVMs
- Feature expansion and kernels
- Two views: maximizing a separating hyperplane margin vs. loss optimization (norm minimization)
- Non-separability & slack
- Sub-gradients

Latent Variable Models
- Meaning: what’s a latent variable?
- General problem: latent variables are (often) not labeled (or difficult)
- General algorithm: expectation maximization
- Problem EM optimizes (and what it doesn’t)

Distributions as Graphs
- Directed graphical models: Bayesian networks
- Undirected graphical models: CRFs, MRFs
- Factor graphs
- Message passing: sum-product

Ensembles & RL
- Decision Trees
- Combining approaches: bagging
- Overview of RL
- RL+Neural methods: deep Q-learning