REUMass Amherst 2015 Data Science Bootcamp

Day 2: Machine Learning and Research Methods

Prof. Ben Marlin
marlin@cs.umass.edu

Prof. David Jensen
jensen@cs.umass.edu

Plan for Day 2:

- Introduction to Machine Learning
- Introduction to Research Methods

Introduction to Machine Learning

What is Machine Learning?

Mitchell (1997): "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."

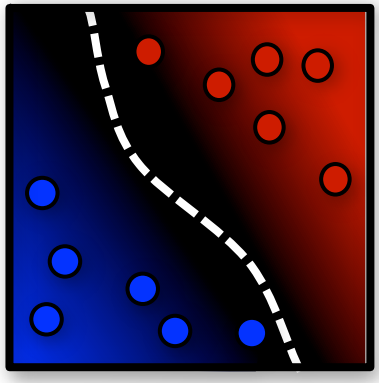
Substitute "data D" for "experience E."

Machine Learning Tasks

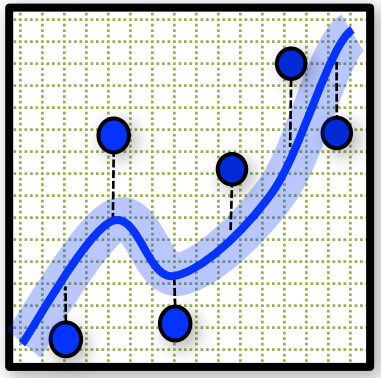
Supervised

Learning to predict.

Classification

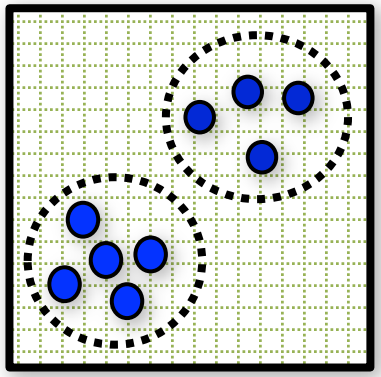


Regression

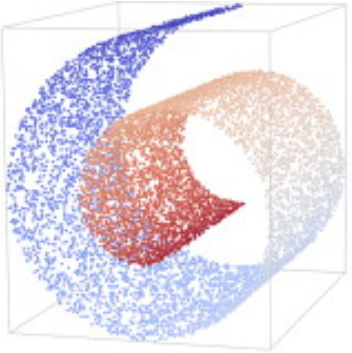


Unsupervised

Learning to organize and represent.



Clustering



Dimensionality Reduction

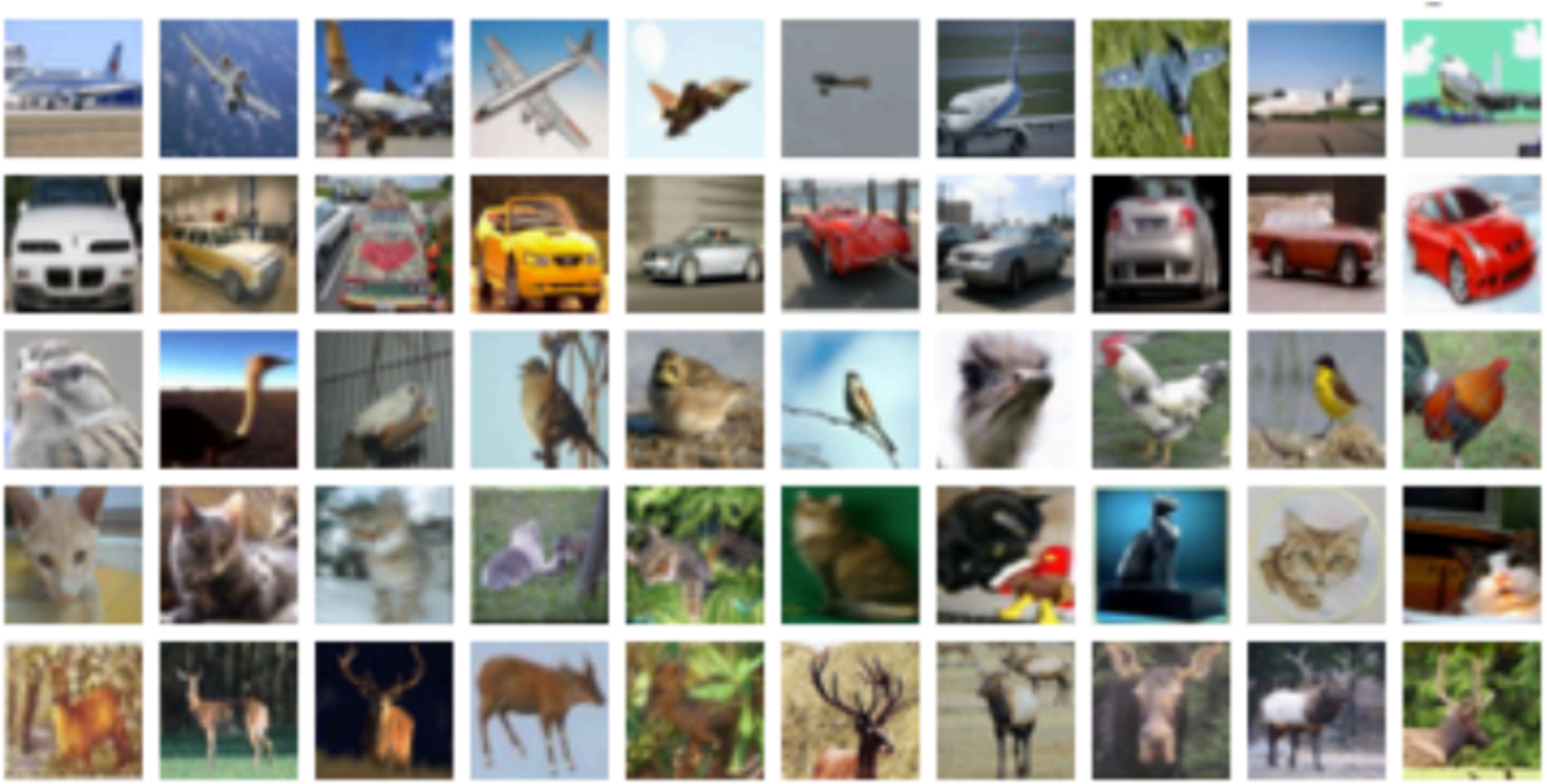
Classification

Definition: The Classification Task

Given a feature vector $\mathbf{x} \in \mathbb{R}^D$ that describes an object that belongs to one of C classes from the set $\mathcal{Y}$, predict which class the object belongs to.

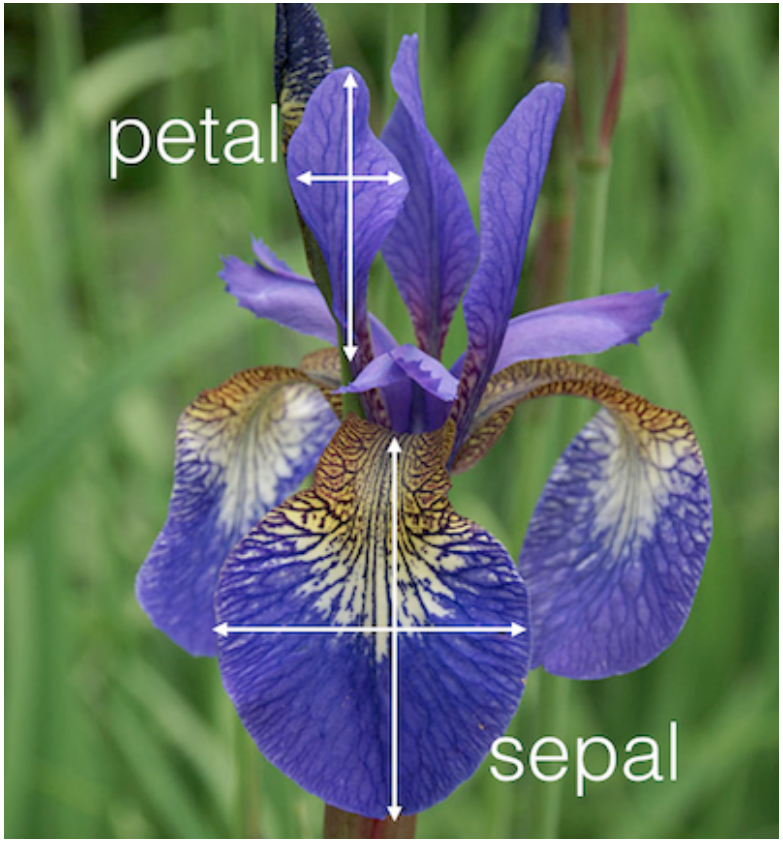
Example: Digits

Example: Natural Images



Example: Fisher's Iris Data

sepal length	sepal width	petal length	petal width	Type
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
⋮	⋮	⋮	⋮	⋮



Classifier Learning

Definition: Classifier Learning

Given a data set of example pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1 : N\}$ where $\mathbf{x}_i \in \mathbb{R}^D$ is a feature vector and $y_i \in \mathcal{Y}$ is a class label, learn a function $f : \mathbb{R}^D \rightarrow \mathcal{Y}$ that accurately predicts the class label y for any feature vector $\mathbf{x}$.

	sepal length	sepal width	petal length	petal width	Type	
x_1	5.1	3.5	1.4	0.2	Setosa	y_1
x_2	4.9	3.0	1.4	0.2	Setosa	y_2
	4.7	3.2	1.3	0.2	Setosa	
	4.6	3.1	1.5	0.2	Setosa	
	⋮	⋮	⋮	⋮	⋮	

Classifier Learning

Definition: Classifier Learning

Given a data set of example pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1 : N\}$ where $\mathbf{x}_i \in \mathbb{R}^D$ is a feature vector and $y_i \in \mathcal{Y}$ is a class label, learn a function $f : \mathbb{R}^D \rightarrow \mathcal{Y}$ that accurately predicts the class label y for any feature vector $\mathbf{x}$.

$f($ 5.1 3.5 1.4 0.2 $) =$ Setosa

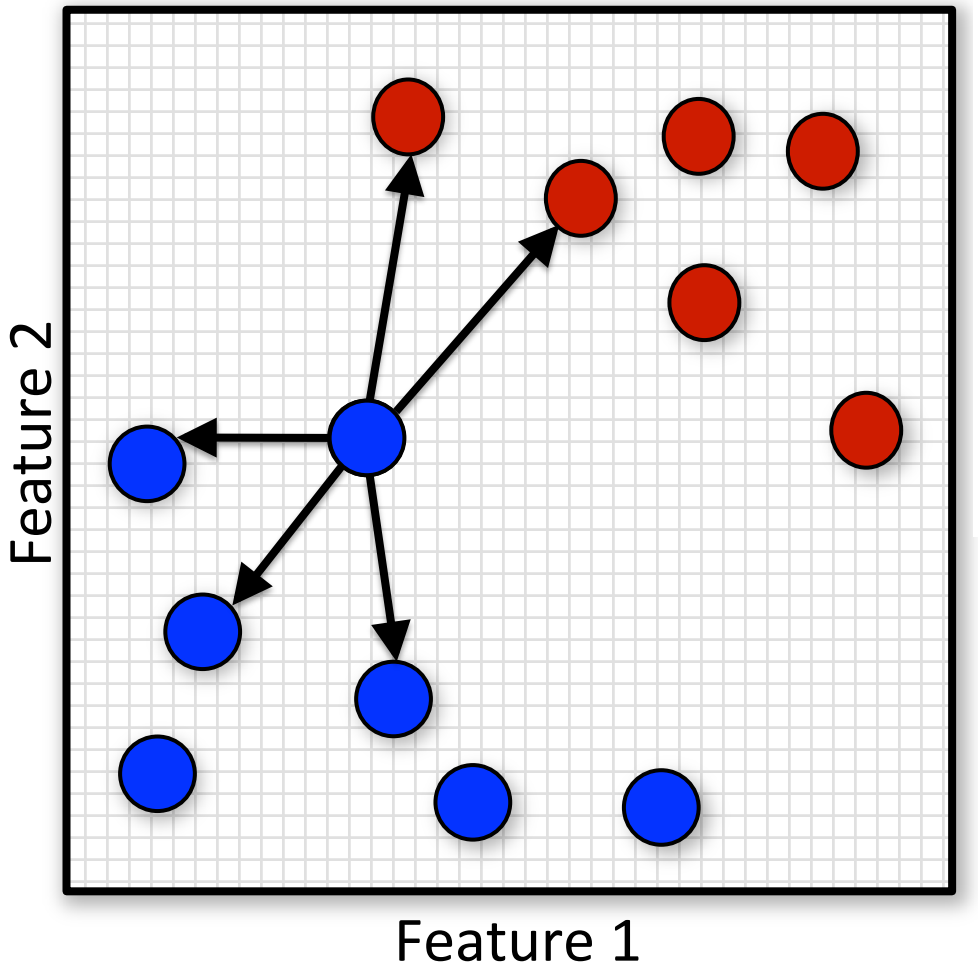
Classifier Performance

Definition: Classification Error Rate

Given a data set of example pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1 : N\}$ and a function $f : \mathbb{R}^D \rightarrow \mathcal{Y}$, the classification error rate of f on $\mathcal{D}$ is:

$$Err(f, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i \neq f(\mathbf{x}_i))$$

K Nearest Neighbors



$$f_{KNN}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} \sum_{i \in \mathcal{N}_K(\mathbf{x})} \mathbb{I}[y_i = y]$$

$$d_p(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_p$$

$$= \left(\sum_{i=1}^{D} |x_d - x'_d|^p \right)^{1/p}$$

- Need to pick number of neighbors K.
- Need to define a distance function.

Exercises: KNN and Fisher's Iris Data

- Q1: What pair of features gives the lowest training error when using K=5?
- Q2: What is the training error rate for K=5 when using all of the features?
- Q3: What value of K gives the lowest error on the training data when using all of the features?