Applied Machine Learning!!!

W207 Section 9
Rasika Bhalerao
rasikabh@berkeley.edu

Aug 23: Welcome!
Nov 8 and 22: No classes

Schedule

Supervised learning methods

	Sync	Topic
2	Aug 30	Linear Regression / Gradient Descent
3	Sep 6	Feature Engineering Bonus: Naive Bayes
4	Sep 13	Logistic Regression
5	Sep 20	Multiclass classification / Eval Metrics Bonus: Reinforcement learning
6	Sep 27	Neural Networks
7	Oct 4	KNN, Decision Trees, Ensembles

Unsupervised learning methods

	Sync	Topic
8	Oct 11	KMeans and PCA Bonus: LDA
9	Oct 18	Text Embeddings Bonus: Language models
10	Oct 25	CNNs Bonus: GANs
11	Nov 1	EDA, Real data, Baselines
12	Nov 15	Fairness / Ethics
13	Nov 29	Fancy Neural Networks
14	Dec 6	Final Presentations

Assignment Schedule

Due Date	Assignment
Aug 28	HW1
Sep 4	HW2
Sep 11	HW3
Sep 18	HW4
Sep 25	HW5
Oct 2	HW6
Oct 16	Group project baseline
Oct 23	HW8
Nov 6	HW9
Nov 20	HW10
Dec 4	Final project notebook + presentation

Behavior expectations

- Healthy disagreement is expected
- Be mindful of one another’s schedules
- Be a good listener
- Have fun in a professional manner
- Share related real-world experience
- Ask questions when something is confusing
- Keep it 100 but be respectful
- Be open-minded to new ideas in the real world and when coding
- On time for group meetings

How are final projects going?

Guidelines:
https://docs.google.com/document/d/1R7mIHOtYXKU8vEQzw10uofb_iK3sgimw8iZLWSTzdgg/edit?usp=sharing

KMeans Clustering

KMeans: 2 sentence overview of the algorithm?

When do we stop iterating?

Why KMeans? Any real world examples?

Why does standardizing features help?

What happens if there is an outlier?

How many ways can you come up with to initialize cluster centers?

Comparing clustering methods: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py

GMMs! Supervised or unsupervised?
What do they actually *do*?

Question

Gaussian mixture models are a probabilistic extension of _______.

GMM demo: https://lukapopijac.github.io/gaussian-mixture-model/

What happens if you add too many Gaussians?

When do we stop iterating?

PCA

What is dimensionality reduction? Is PCA feature selection?

Why do we use PCA?

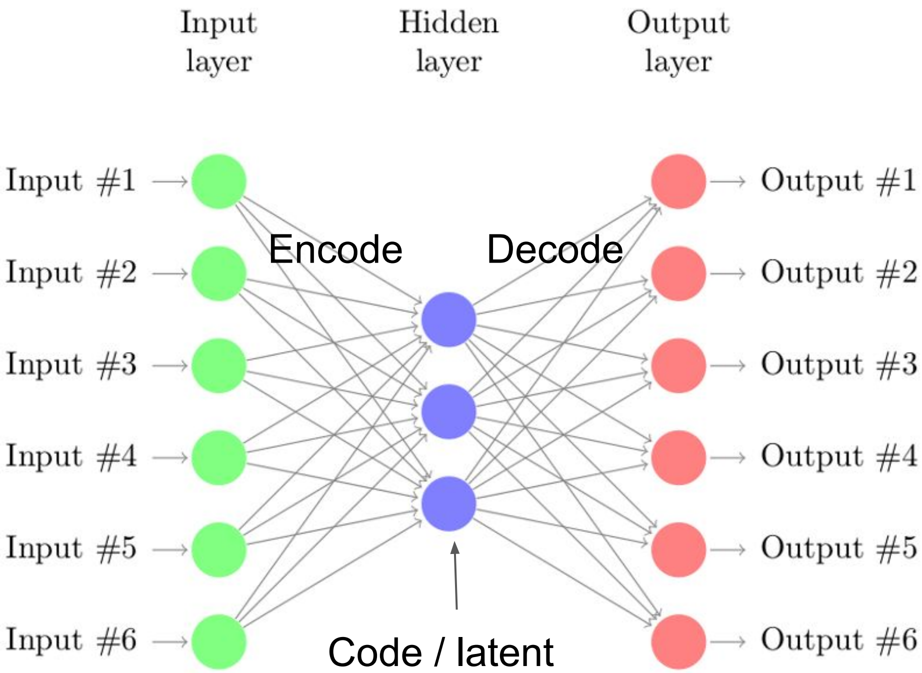
Why do we use PCA?

- Less computation
- Sparse data - “curse of dimensionality”
- Visualization
- Reduce noise
- Prevent overfitting
- Find most useful component
- Data compression
- Uncovering hidden structure

Does it make sense to do PCA on the features for decision trees? Logistic regression?

How do we apply PCA on the train, dev, and test sets differently?

Auto-encoders

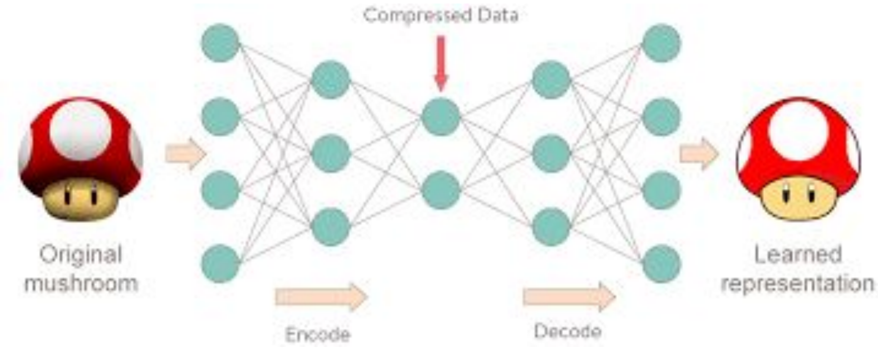


Encode

Decode

Code / latent representation

Y = X



https://iq.opengenus.org/types-of-autoencoder/

View options (at the top) → Annotate

Async Practice Quiz Questions (vote!)

If Vectors A and B have unit length, their dot product is the angle between them.	True	False
Single linkage refers to the method of linking two clusters based on their most distant items.	True	False
The K-means algorithm selects an appropriate number of clusters.	True	False
K-means for color quantization maps RGB values to a small set of representative colors.	True	False
Estimating all the parameters of a multivariate normal via maximum likelihood can be done in just two passes through the data.	True	False
The EM algorithm with soft assignment means that every data point is assigned a single Gaussian component in each expectation step.	True	False
The SVD can be used to approximately reconstruct a data matrix M from a small set of (transformed) features.	True	False

Notebook!

To access later:
https://github.com/MIDS-W207/rasikabh/blob/main/live_sessions/Week8.ipynb

Also, if you want last semester’s assignments:
https://github.com/MIDS-W207/coursework