Possible Practical Machine Learning Project Ideas:

Your individual project is an opportunity for you to explore an interesting machine learning problem of your choice in the context of a real-world data set. Below, you will find some project ideas both practical and theoretical machine learning as example. But the best idea would be to combine machine learning with problems in your own research area. Your individual project must be about new things you have done this semester; you can't use results you have developed in the past.

1. Recognition of character recognition (digits) data

   Optical character recognition, and the simpler digit recognition task, has been the focus of much ML research. We have three datasets on this topic. The first tackles the more general OCR task, on a small vocabulary of words: (Note that the first letter of each word was removed, since these were capital letters that would make the task harder for you.) The task here is to learn a classifier to recognize the letter/digit and then apply a clustering/dimensionality reduction algorithm on this data, see if you get better classification on this lower dimensional space. You are required to rest at two classification algorithms on the following two datasets.

   ○ **OCR:** **http://ai.stanford.edu/~btaskar/ocr/**
   ○ **MNIST:** **http://yann.lecun.com/exdb/mnist/**

2. **Imbalanced Data Classification**

   Many important machine learning tasks involve classifying imbalanced datasets where one class has much more data points than other classes. Such imbalanced datasets appear in anomaly detection, medical diagnosis, information retrieval and rare event detection (disease outbreak detection). There are many approaches to address the imbalanced classification such as cost-sensitive approach, sampling methods and direct approach of optimization certain performance measures.

   References:

Handling imbalanced datasets: A review

https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/

Stochastic Proximal AUC Maximization

Deep learning for imbalanced data

3. Multi-class Classification from UCI Repository

We have addressed various methods on binary classification using SVM. In practice, many real-life applications have classes more than 3. The following link is a collection of various datasets for multi-class classification.

https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html

Suggested project is to test at least two multi-class SVM methods for at least 5 multi-class datasets in the above link.

References:

C. Hsu and C. Lin. A Comparison of Methods for Multi-class Support Vector Machines.

4. Multi-task Learning

School dataset (See the blackboard)

The School dataset is from the Inner London Education Authority. It consists of the examination scores of 15,362 students from 139 secondary schools in 1985, 1986 and 1987. There are 139 tasks in total, corresponding to examination scores prediction in each school. The input features include the year of the examination, 4 school dependent features and 3 student-dependent features

Suggest project: to examine two or three multi-task learning methods on the school datasets to discuss their performance (generalization and computationally performance).

References:

Multi-Task Feature Learning.

A framework for learning predictive structures from multiple tasks and unlabeled data

Multi-task Learning

A spectral regularization framework for multi-task structure learning

5. **Metric Learning:**

- Kilian Q. Weinberger and Lawrence K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. Journal of Machine Learning Research (JMLR), 10: 207–244, 2009.
- Yiming Ying and Peng Li. Distance Metric Learning with Eigenvalue Optimization. Journal of Machine Learning Research (JMLR), 13:1–26, 2012.
- Yiming Ying, Kaizhu Huang, and Colin Campbell. Sparse Metric Learning via Smooth Optimization. In Advances in Neural Information Processing Systems (NIPS) 22, pages 2214–2222, 2009.
- Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart J. Russell. Distance Metric Learning with Application to Clustering with Side-Information. In Advances in Neural Information Processing Systems (NIPS) 15, pages 505–512, 2002.

  You can find more related work on metric learning https://www.cs.cmu.edu/~liuy/dist_overview.pdf

6. Unconstrained Face Verification

Recently, considerable research efforts are devoted to the unconstrained face verification problem, the task of which is to predict whether two face images represent the same person or not. The face images are taken under unconstrained conditions and show significant variations in complex background, lighting, pose, and expression (see e.g. Figure 1). In addition, the evaluation procedure for face verification typically assumes that the person identities in the training and test sets are exclusive, requiring the prediction of never-seen-before faces. Both factors make face verification very challenging, see e.g.

Labeled Faces in the Wild (LFW) dataset.

Please read the following papers carefully on how to use the datasets:

○ Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. **Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.**

The task is to construct a classifier to discriminate the similar pairs from the same person (class 1) from dissimilar pairs from different persons (class -1). One can use SVM and metric learning to construct such classifiers.

More related work:

○ Q. Cao, Y. Ying and Peng Li, Similarity metric learning for face recognition. *ICCV*, 2013.

- Junlin Hu, Jiwen Lu, and Yap-Peng Tan.
  Discriminative Deep Metric Learning for Face Verification in the Wild.
  CVPR, 2014.
- Tal Hassner, Shai Harel, Eran Paz and Roee Enbar.
  Effective Face Frontalization in Unconstrained Images. *CVPR*, 2015.
- Karen Simonyan, Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman.
  Fisher Vector Faces in the Wild. *BMVC*, 2013.

More related work can be found in
Erik Learned-Miller, Gary B. Huang et~al. **Labeled Faces in the Wild: A Survey**

7. Solar radiation forecasting

Forecasting the output power of solar systems is required for the good operation of the power grid or for the optimal management of the energy fluxes occurring into the solar system. It is essential to focus the prediction on the solar irradiance. The global solar radiation forecasting can be performed by machine learning algorithms based on meteorological data . In this context, one can use neural networks or support vector regression and other robust regression models.

The time series nature of the solar radiation data makes it different from normal non-temporal data. Below is a paper on how to deal with prediction with time series (sequential) data.

Thomas G. Dietterich. Machine Learning for Sequential Data: A Review

The project should properly process the solar radiation data and introduce proper evaluation methods of different machine learning algorithms.

References: Anish Agarwal et.al. **Time Series Forecasting via Matrix Estimation**

Cyril Voyant et~al. Machine learning methods for solar radiation forecasting: A review. Renewable Energy, Vol 15: 569-582 (2017).

Send me an email for the data set.

8. Wildfire Prediction

The goal of the project is to design a novel approach to the problem of spatial and temporal modeling of fires and evaluating the importance of socioeconomic and anthropogenic factors for the region of south-east Siberia. The spatial information from this region will include weather, topography, and wildfire data as well as human factors. The models will be designed using the machine learning techniques random forest (RF) and maximum entropy (MAXENT) as well as designing a generalized linear model (GLM) as a benchmark for comparison. The standard metrics for each model (deviance, gain, and mean square error) will be used to measure the model's individual performance. The AUC score will be used to compare the models produced by the different methods. Model predictions produced by each model will be compared by using prediction maps and Spearman's correlation coefficient between pairs of maps.

9. **Seizure Detection**

Epilepsy is a devastating neurological disorder that affects approximately 65 million people worldwide. Electroencephalography (EEG) recording has become an essential tool in evaluating seizure activity, critical for both epilepsy research and treatment. However, due to the randomly occurring and infrequent nature of seizures, seizure evaluation requires continuous, long-term EEG monitoring, which produces a large volume data. Analysis of this EEG data is extremely
burdensome and labor-intensive, often involving manual visual scanning of large EEG records
by trained personnel.

The aim of this project is to develop a novel seizure detection algorithm
with low computational cost and high accuracy using machine learning.

Talk to me about datasets and references

10. Rare event classification

- Extreme Rare Event Classification using Autoencoders in Keras
- Ranjan, C., Mustonen, M., Paynabar, K., & Pourak, K. (2018). Dataset: Rare Event Classification in Multivariate Time Series. *arXiv preprint arXiv:1809.10717*
- Time-series forecasting with deep learning & LSTM autoencoders
- Complete code: LSTM Autoencoder

11. Face recognition data

There are two data sets for this problem. The first dataset contains 640 images of faces. The faces themselves are images of 20 former Machine Learning students and instructors, with about 32 images of each person. Images vary by the pose (direction the person is looking), expression (happy/sad), face jewelry (sun glasses or not), etc. This gives you a chance to consider a variety of classification problems ranging from person identification to sunglass detection. The data, documentation, and associated code are available here:
CMU Machine Learning Faces
Available Software: The same website provides an implementation of a neural network classifier for this image data.

The second data set consists of 2253 female and 1745 male rectified frontal face images scraped from the hotornot.com website by Ryan White along with user ratings of attractiveness. The data set can be found here: **Facial Attractiveness Images.**

Try SVM's on this data, and compare their performance to that of the provided neural networks and apply a clustering algorithm to find "similar" faces. For learning a facial attractiveness classifier. A recent NIPS paper on the topic of predicting facial attractiveness can be found here.

12. Image Segmentation Dataset

The goal is to segment images in a meaningful way. Berkeley collected three hundred images and paid students to hand-segment each one (usually each image has multiple hand-segmentations). Two-hundred of these images are training images, and the remaining 100 are test images. The dataset includes code for reading the images and ground-truth labels, computing the benchmark scores, and some other utility functions. It also includes code for a segmentation example. This dataset is new and the problem unsolved, so there is a chance that you could come up with the leading algorithm for your project.
http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

Possible method 1: Region-Based Segmentation
Most segmentation algorithms have focused on segmentation based on edges or based on discontinuity of color and texture. The ground-truth in this dataset, however, allows supervised learning algorithms to segment the images based on statistics calculated over regions. One way to do this is to "oversegment" the image into superpixels (Felzenszwalb 2004, code available) and merge the superpixels into larger segments. Come up with a set of features to represent the superpixels (probably based on color and texture), a classifier/regression algorithm (suggestion: boosted decision trees) that allows you to estimate the likelihood that two superpixels are in the same segment, and an algorithm for

segmentation based on those pairwise likelihoods. Since this project idea is fairly time-consuming focusing on a specific part of the project may also be acceptable. **Papers to read:** Some segmentation papers from Berkeley are available here

Possible method 2: Supervised vs. Unsupervised Segmentation Methods
Write two segmentation algorithms (these may be simpler than the one above): a supervised method (such as logistic regression) and an unsupervised method (such as K-means). Compare the results of the two algorithms. For your write-up, describe the two classification methods that you plan to use.
Papers to read: Some segmentation papers from Berkeley are available here

13. **Brain imaging data (fMRI)** This data is available here

This data set contains a time series of images of brain activation, measured using fMRI, with one image every 500 msec. During this time, human subjects performed 40 trials of a sentence-picture comparison task (reading a sentence, observing a picture, and determining whether the sentence correctly described the picture). Each of the 40 trials lasts approximately 30 seconds. Each image contains approximately 5,000 voxels (3D pixels), across a large portion of the brain. Data is available for 12 different human subjects.

SVMs have been used with this data to predict when the subject was reading a sentence versus perceiving a picture. Both of these classify 8-second windows of data into these two classes, achieving around 85% classification accuracy [Mitchell et al, 2004]. In the middle term, you should have run at least one classification algorithm on this data and measured its accuracy using a cross validation test. This will put you in a good position to explore refinements of the algorithm, alternative feature encodings for the data, or competing algorithms, by the end of the semester.

Explore the use of dimensionality-reduction methods to improve classification accuracy with this data. Given the extremely high dimension of the input (5000 voxels times 8 images) to the classifier, it is sensible to explore methods for reducing this to a small number of dimension. For example, consider PCA, hidden layers of neural nets, or other relevant dimensionality reducing methods.

Papers to read: T. Mitchell (2004). Learning to Decode Cognitive States from Brain Images. Machine Learning, 57, 145–175, 2004