Given the schedule (weekly topics) define the Learning Outcomes and assessments

|  |  |
| --- | --- |
| **Weekly Schedule** | |
| **Week** | **Topics** |
| 1 | Overview of Machine Learning: history, relation to classical statistics **Text:** Chapter 1; |
| 2 | **Linear Regression:** model specification issues, maximum likelihood estimation, ridge regression **Text:** Chapter 7, except 7.4 and 7.6 |
| 3 | **Logistic Regression:** model fitting methods (steepest descent, Newton’s, L1 and L2 regularization, etc.) **Online learning algorithms**: structured prediction, regret minimization, stochastic optimization, risk minimization, LMS algorithm, perceptron algorithm, relation to Bayesian techniques. **Text:** Chapter 8.1-8.3.6, 8.5, 13.3, |
| 4 | **Support Vector Machines:** uses in regression and classification, optimization, choice of parameter C, probabilistic interpretation **Kernel Methods:** RBF kernels, Mercer kernels, Matern kernels, Linear kernels, String kernels, Kernels derived from probabilistic generative models; the “kernel trick” and its uses: nearest neighbor classification, K-medoids clustering, ridge regression, and PCA **Text:** Chapter 14.1, 14.2, 14.5, supplemental notes by Ng |
| 5 | **Decision Trees:** basis in information theory, when appropriate to use, growing and pruning, detecting and avoiding over-fitting; pros and cons with respect to other ML techniques; random forests; ID3 and related algorithms; use of bias; training with incomplete data **Boosting:** why it works so well, specific types of boosting, ensemble learning **Text:** Chapters 2.8, 16.2, 16.4, 16.6 , supplemental notes by Mitchell, Yoav & Schapire |
| 6 | **Deep Learning:** background in feed-forward neural networks (multilayer perceptrons) and Restricted Boltzmann machines; deep generative models (sigmoid networks, Boltzmann machines, belief networks) **Text:** Chapters 16.5, 27.7, 28.1-28.2 |
| 7 | **Deep Learning (continued):** training of deep networks, applications of deep networks **Midterm Review** **Text:** Chapters 28.3-28.4 |
| 8 | **Midterm Exam Clustering:** measures of dissimilarity; k-means clustering, clustering in mixture models **Text:** Chapters 25.1, 11.1-11.3 |
| 9 | **Expectation Maximization:** basic idea, Jensen’s inequality, the EM algorithm, EM applied to various distributions (mixture of Gaussians), theoretical basis **Text:** Chapters 11.4, supplemental notes by Ng |
| 10 | **Dimensionality Reduction:** classical Principal Components Analysis, PCA theorem; Singular Value Decomposition, Probabilistic PCA, EM algorithm for PCA, PCA for categorical data, PCA for paired and multi-view data **Text:** Chapters 12.2, 12.4, 12.5 |
| 11 | **Graphical Techniques:** Bayesian nets, Directed Gaussian Models (DGM), plate notation, d-separation, Markov random fields, Hammersley-Clifford theorem and its effects, training maxent models, pseudo-likelihood, stochastic maximum likelihood learning **Text:** Chapter 10, 19.1-19.5 |

COM-3920: Machine Learning

Learning Objectives

Course Level Learning Outcomes

Use and analyze existing learning algorithms, including methods for classification, regression, structured prediction, clustering, and representation learning

Integrate multiple facets of practical machine learning in a single system: data preprocessing, learning, regularization and model selection

Describe the formal properties of models and algorithms for learning and explain the practical implications of those results

Compare and contrast different paradigms for learning (supervised, unsupervised, etc.)

**Week 1**.

**Topic:** Overview of Machine Learning.

**Outcomes**

* Formulate a well-posed learning problem for a real-world task by identifying the task, performance measure, and training experience
* Describe common learning paradigms in terms of the type of data available and when, the form of prediction, and the structure of the output prediction
* Identify examples of the ethical responsibilities of an ML expert

**Questions**

* What Does it Mean to Learn?
  + *The general supervised approach to machine learning: a learning algorithm reads in training data and computes a learned function f.*
* What is the difference between memorization and generalization?

**Evidence**

* Ask students to formulate a learning problem (in class)
* Homework (focusing on revisions and assessing what they know in terms of basic probability and statistics, calculus, and linear algebra)

**Week 2**.

**Topic: Linear Regression**

* + To know simple linear regression, intercept slope coefficient parameter, least squares, sum of squares, population regression line, bias and unbiased, standard error, residual standard error
  + Design k-NN Regression and Decision Tree Regression
  + To Understand how to estimate the regression coefficients
  + Implement learning for Linear Regression using three optimization techniques: (1) closed form, (2) gradient descent, (3) stochastic gradient descent
  + To be able to choose a Linear Regression optimization technique that is appropriate for a particular dataset by analyzing the tradeoff of computational complexity vs. convergence speed

**Questions**

* How does KNN work and what are the pros and cons of KNN? How can we tune the hyperparameter K?
* What are the differences between the KNN classifier and KNN regression methods?
* How do we choose the step-size and convergence criteria?
* How do we asses model accuracy?
* Is There a Relationship Between the Response and Predictors?

**Evidence**

* Quiz on KNN and Regression
* Case study: Predicting Housing Prices
* Homework
  + Problems
  + Applying concepts on a dataset (Python)

Topic: Optimization for ML (Linear Regression)

* + Apply gradient descent to optimize a function
  + Apply stochastic gradient descent (SGD) to optimize a function
  + Apply knowledge of zero derivatives to identify a closed-form solution (if one exists) to an optimization problem
  + Distinguish between convex, concave, and nonconvex functions
  + Obtain the gradient (and Hessian) of a (twice) differentiable function

**Evidence**

Quiz: Linear Regression with Multiple Variables (Andrew Ng, Coursera)

Generative Models

1. Sampling, Generative vs. Discriminative
   1. Sample from common probability distributions
   2. Write a generative story for a generative or discriminative classification or regression model
   3. Provide a probabilistic interpretation of linear regression
   4. Use the chain rule of probability to contrast generative vs. discriminative modeling
   5. Define maximum likelihood estimation (MLE) and maximum conditional likelihood estimation (MCLE)

**Evidence**

* Homework (focusing on revisions and assessing what they know in terms of basic probability and statistics, calculus, and linear algebra)

1. MLE and MAP
   1. Recall probability basics, including but not limited to: discrete and continuous random variables, probability mass functions, probability density functions, events vs. random variables, expectation and variance, joint probability distributions, marginal probabilities, conditional probabilities, independence, conditional independence
   2. Describe common probability distributions such as the Beta, Dirichlet, Multinomial, Categorical, Gaussian, Exponential, etc.
   3. State the principle of maximum likelihood estimation and explain what it tries to accomplish
   4. State the principle of maximum a posteriori estimation and explain why we use it
   5. Derive the MLE or MAP parameters of a simple model in closed form
2. Naive Bayes
   1. Write the generative story for Naive Bayes
   2. Create a new Naive Bayes classifier using your favorite probability distribution as the event model
   3. Apply the principle of maximum likelihood estimation (MLE) to learn the parameters of Bernoulli Naive Bayes
   4. Motivate the need for MAP estimation through the deficiencies of MLE
   5. Apply the principle of maximum a posteriori (MAP) estimation to learn the parameters of Bernoulli Naive Bayes
   6. Select a suitable prior for a model parameter
   7. Describe the tradeoffs of generative vs. discriminative models
   8. Implement Bernoulli Naives Bayes
   9. Employ the method of Lagrange multipliers to find the MLE parameters of Multinomial Naive Bayes
   10. Describe how the variance affects whether a Gaussian Naive Bayes model will have a linear or nonlinear decision boundary

**Questions**

* What are the characteristics of different important distributions: normal, t, beta; what is their mean, comments on skewness, kurtosis.
* Examples of generative models (e.g., standard regression model) and discriminative (e.g., Naive Bayes)
* Find the MLE for the mean of normally distributed variables (intro example) and (more applied) MLE for regression beta coefficient.

**Evidence**

* How do you recognize what distribution a sample comes from? Normal, t, beta.
  + *The students are split in pairs: one generates a sample from a distribution, the other has to use graphs (histogram,q-q plot), hypothesis tests (goodness of fit test) to figure out which distribution the generated sample comes from.*
* Use a Naïve Bayes model to predict Sports vs Non Sports messages.

*(adapt it from here: https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/)*

**Week 3**.

Logistic Regression (Probabilistic Learning)

* 1. Apply the principle of maximum likelihood estimation (MLE) to learn the parameters of a probabilistic model
  2. Given a discriminative probabilistic model, derive the conditional log-likelihood, its gradient, and the corresponding Bayes Classifier
  3. Explain the practical reasons why we work with the **log** of the likelihood
  4. Implement logistic regression for binary or multiclass classification
  5. Prove that the decision boundary of binary logistic regression is linear
  6. For linear regression, show that the parameters which minimize squared error are equivalent to those that maximize conditional likelihood

k-Nearest Neighbors

* 1. Describe a dataset as points in a high dimensional space [CIML]
  2. Implement k-Nearest Neighbors with O(N) prediction
  3. Describe the inductive bias of a k-NN classifier and relate it to feature scale [a la. CIML]
  4. Sketch the decision boundary for a learning algorithm (compare k-NN and DT)
  5. State Cover & Hart (1967)'s large sample analysis of a nearest neighbor classifier
  6. Invent "new" k-NN learning algorithms capable of dealing with even k
  7. Explain computational and geometric examples of the curse of dimensionality

Model Selection

* 1. Plan an experiment that uses training, validation, and test datasets to predict the performance of a classifier on unseen data (without cheating)
  2. Explain the difference between (1) training error, (2) validation error, (3) cross-validation error, (4) test error, and (5) true error
  3. For a given learning technique, identify the model, learning algorithm, parameters, and hyperparamters
  4. Define "instance-based learning" or "nonparametric methods"
  5. Select an appropriate algorithm for optimizing (aka. learning) hyperparameters

Perceptron

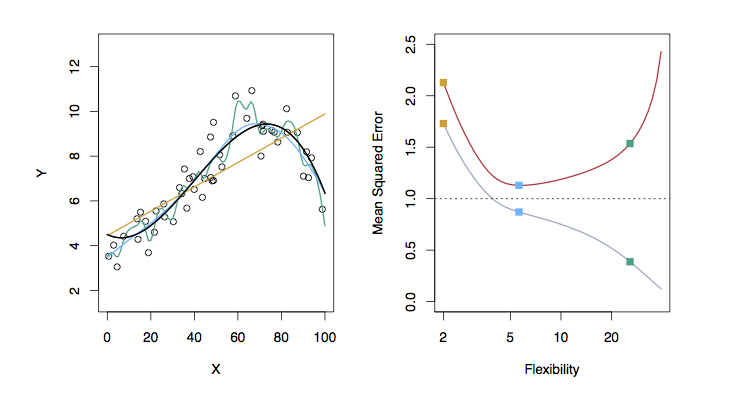
* 1. Explain the difference between online learning and batch learning
  2. Implement the perceptron algorithm for binary classification [CIML]
  3. Determine whether the perceptron algorithm will converge based on properties of the dataset, and the limitations of the convergence guarantees
  4. Describe the inductive bias of perceptron and the limitations of linear models
  5. Draw the decision boundary of a linear model
  6. Identify whether a dataset is linearly separable or not
  7. Defend the use of a bias term in perceptron (shifting points after projection onto weight vector)

Questions:

1. How do you interpret the odds ratios of logistic regression?

(good explanation here: <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/>)

1. Explain difference between train and test error



LEFT

Black: Truth

Orange: Linear Estimate

Blue: smoothing spline

Green: smoothing spline (more flexible)

RIGHT

RED: Test MES

Grey: Training MSE

Dashed: Minimum possible test MSE (irreducible error)

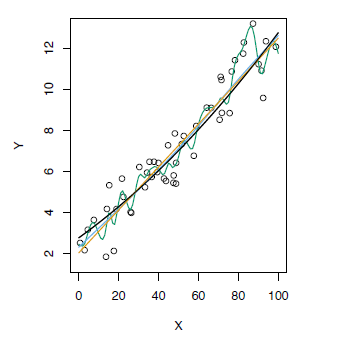
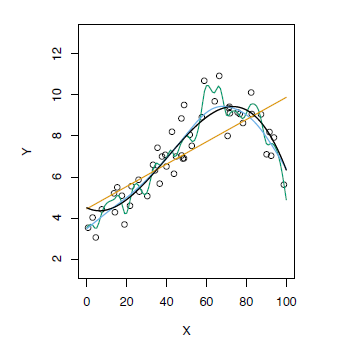
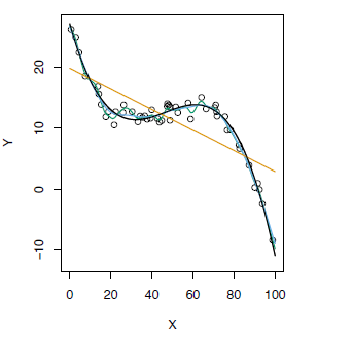
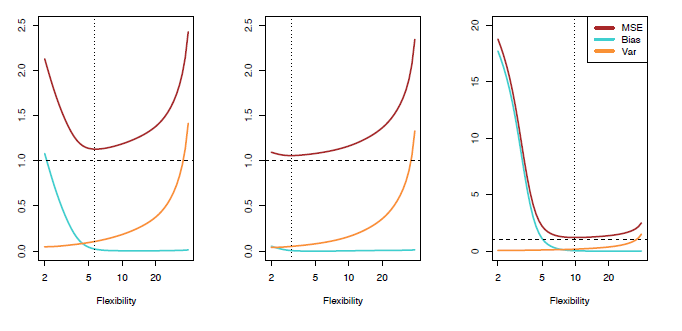
RIGHT

RED: Test MES

Grey: Training MSE

Dashed: Minimum possible test MSE (irreducible error)

1. Explain difference between bias and variance

**Evidence:**

* Implement several spam detection logistic regression models and use cross validation to compare between them https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522
* Ask students to build a simple neural network using tensorflow. <https://www.tensorflow.org/tutorials/>

**Week 4.**

SVMs

* 1. Motivate the learning of a decision boundary with large margin
  2. Compare the decision boundary learned by SVM with that of Perceptron
  3. Distinguish unconstrained and constrained optimization
  4. Compare linear and quadratic mathematical programs
  5. Derive the hard-margin SVM primal formulation
  6. Derive the Lagrangian dual for a hard-margin SVM
  7. Describe the mathematical properties of support vectors and provide an intuitive explanation of their role
  8. Draw a picture of the weight vector, bias, decision boundary, training examples, support vectors, and margin of an SVM
  9. Employ slack variables to obtain the soft-margin SVM
  10. Implement an SVM learner using a black-box quadratic programming (QP) solver

Kernels

* 1. Employ the kernel trick in common learning algorithms
  2. Explain why the use of a kernel produces only an implicit representation of the transformed feature space
  3. Use the "kernel trick" to obtain a computational complexity advantage over explicit feature transformation
  4. Sketch the decision boundaries of a linear classifier with an RBF kernel

K-Means

1. Distinguish between coordinate descent and block coordinate descent
2. Define an objective function that gives rise to a "good" clustering
3. Apply block coordinate descent to an objective function preferring each point to be close to its nearest objective function to obtain the K-Means algorithm
4. Implement the K-Means algorithm
5. Connect the nonconvexity of the K-Means objective function with the (possibly) poor performance of random initialization

PCA and Dimensionality Reduction

1. Define the sample mean, sample variance, and sample covariance of a vector-valued dataset
2. Identify examples of high dimensional data and common use cases for dimensionality reduction
3. Draw the principal components of a given toy dataset
4. Establish the equivalence of minimization of reconstruction error with maximization of variance
5. Given a set of principal components, project from high to low dimensional space and do the reverse to produce a reconstruction
6. Explain the connection between PCA, eigenvectors, eigenvalues, and covariance matrix
7. Use common methods in linear algebra to obtain the principal components

**Week 5**

Decision Trees

1. Formalize a learning problem by identifying the input space, output space, hypothesis space, and target function
2. Implement Decision Tree training and prediction
3. Use effective splitting criteria for Decision Trees and be able to define entropy, conditional entropy, and mutual information / information gain
4. Explain the difference between memorization and generalization [CIML]
5. Describe the inductive bias of a decision tree
6. Judge whether a decision tree is "underfitting" or "overfitting"
7. Explain the difference between true error and training error
8. Implement a pruning or early stopping method to combat overfitting in Decision Tree learning

Feature Engineering / Regularization

* 1. Engineer appropriate features for a new task
  2. Use feature selection techniques to identify and remove irrelevant features
  3. Identify when a model is overfitting
  4. Add a regularizer to an existing objective in order to combat overfitting
  5. Explain why we should **not** regularize the bias term
  6. Convert linearly inseparable dataset to a linearly separable dataset in higher dimensions
  7. Describe feature engineering in common application areas

**Weeks 6 to 10**

Deep Learning

1. Neural Networks
   1. Explain the biological motivations for a neural network
   2. Combine simpler models (e.g. linear regression, binary logistic regression, multinomial logistic regression) as components to build up feed-forward neural network architectures
   3. Explain the reasons why a neural network can model nonlinear decision boundaries for classification
   4. Compare and contrast feature engineering with learning features
   5. Identify (some of) the options available when designing the architecture of a neural network
   6. Implement a feed-forward neural network
2. Backpropagation / Deep Learning
   1. Construct a computation graph for a function as specified by an algorithm
   2. Carry out the backpropagation on an arbitrary computation graph
   3. Construct a computation graph for a neural network, identifying all the given and intermediate quantities that are relevant
   4. Instantiate the backpropagation algorithm for a neural network
   5. Instantiate an optimization method (e.g. SGD) and a regularizer (e.g. L2) when the parameters of a model are comprised of several matrices corresponding to different layers of a neural network
   6. Apply the empirical risk minimization framework to learn a neural network
   7. Use the finite difference method to evaluate the gradient of a function
   8. Identify when the gradient of a function can be computed at all and when it can be computed efficiently

**Week 11 and 12**

Graphical Models

1. Hidden Markov Models
   1. Show that structured prediction problems yield high-computation inference problems
   2. Define the first order Markov assumption
   3. Draw a Finite State Machine depicting a first order Markov assumption
   4. Derive the MLE parameters of an HMM
   5. Define the three key problems for an HMM: evaluation, decoding, and marginal computation
   6. Derive a dynamic programming algorithm for computing the marginal probabilities of an HMM
   7. Interpret the forward-backward algorithm as a message passing algorithm
   8. Implement supervised learning for an HMM
   9. Implement the forward-backward algorithm for an HMM
   10. Implement the Viterbi algorithm for an HMM
   11. Implement a minimum Bayes risk decoder with Hamming loss for an HMM
2. Bayesian Networks
   1. Identify the conditional independence assumptions given by a generative story or a specification of a joint distribution
   2. Draw a Bayesian network given a set of conditional independence assumptions
   3. Define the joint distribution specified by a Bayesian network
   4. User domain knowledge to construct a (simple) Bayesian network for a real-world modeling problem
   5. Depict familiar models as Bayesian networks
   6. Use d-separation to prove the existence of conditional independencies in a Bayesian network
3. Employ a Markov blanket to identify conditional independence assumptions of a graphical model
4. Develop a supervised learning algorithm for a Bayesian network
5. Use samples from a joint distribution to compute marginal probabilities
6. Sample from the joint distribution specified by a generative story
7. Implement a Gibbs sampler for a Bayesian network

**Week 13**

Reinforcement Learning

1. Reinforcement Learning: Value & Policy Iteration
   1. Compare the reinforcement learning paradigm to other learning paradigms
   2. Cast a real-world problem as a Markov Decision Process
   3. Depict the exploration vs. exploitation tradeoff via MDP examples
   4. Explain how to solve a system of equations using fixed point iteration
   5. Define the Bellman Equations
   6. Show how to compute the optimal policy in terms of the optimal value function
   7. Explain the relationship between a value function mapping states to expected rewards and a value function mapping state-action pairs to expected rewards
   8. Implement value iteration
   9. Implement policy iteration
   10. Contrast the computational complexity and empirical convergence of value iteration vs. policy iteration
   11. Identify the conditions under which the value iteration algorithm will converge to the true value function
   12. Describe properties of the policy iteration algorithm
2. Reinforcement Learning: Q-Learning
   1. Apply Q-Learning to a real-world environment
   2. Implement Q-learning
   3. Identify the conditions under which the Q-learning algorithm will converge to the true value function
   4. Adapt Q-learning to Deep Q-learning by employing a neural network approximation to the Q function
   5. Describe the connection between Deep Q-Learning and regression

References

Several of these