CSE 512: Learning goals, lectures 1-6

Nomenclature

Define / diffrentiate the following terms

- discriminative vs generative model
- model vs loss function vs data
- prior vs posterior vs likelihood distributions
- MLE vs MAP
- expected risk vs empirical risk
- overfitting vs underfitting
- training vs generalization error
- prediction, forecasting
- classification vs regression

Linear models for classification

1. Consider the following tasks and features. Would you propose using a linear classifier (without manipulating the features beyond scaling) to accomplish these tasks? Why or why not?

    - Predicting tomorrow's weather based on last week's weather
    - Predicting whether I will buy a house based on its price, square footage, crime rating
    - Predicting whether I would buy a small house (fit for 3-4 people) based on how many kids I have
    - Predicting which player will win at a game of chess based on the current configuration

    Overall, what are your criteria for using a linear model vs a different model?

2. **Regression** Describe briefly how you can extend linear modeling to nonlinear modeling using bases functions. As an example, consider fitting a signal $x_t \in [-1, 1]$ for $t \in [0, T]$ using bases functions $g_k(t) = \sin(kt)$. Describe your procedure for fitting and forecasting.

3. **Classification** For the generalized linear model in classification

    $$\mathcal{L}(\theta) = \frac{1}{m}\sum_{i=1}^{m} g(y_i \theta^T x_i)$$

    - Evaluate the gradient and Hessian of $\mathcal{L}$ (in terms of g)
    - Discuss what kinds of functions g may be used in binary classification and why
    - Evaluate the Lipschitz constant L and strong convexity constant μ (in terms of properties of X, y, and g)
    - Write out the iteration scheme for gradient descent (including suggesting an appropriate step size) for minimizing this loss function.
    - Describe what may happen if we pick a step size that is bigger than the one you proposed.
    - What value of g is used in logistic regression? Describe the properties of this g, and compare it with that used in exponential and hinge loss implementations.

4. The prediction for a linear classifier looks like

$$y = \mathbf{sign}(x^T\theta)$$

where x is a data vector, and y is its label. Describe the *margin*, what a margin maximizing method is, and how margin maximizing methods work. List at least 2 methods which are margin maximizing, and justify why they have that effect.

Training continuous models

1. Describe the difference between a loss function, a regularization penalty, and a performance metric. Categorize the following into one or more of these:
   - logistic loss
   - least squares deviation
   - mean squared error
   - 2-norm
   - misclassification rate
2. We have shown that sometimes it is advantageous to add a 2-norm squared penalty to loss functions. Give at least 2 different reasons why that might be desirable for
   - linear regression
   - logistic regression
   - other loss functions?
3. For what kind of loss functions would we use gradient descent for training? (convex, nonconvex, concave), (likelihood, utility, loss?)
4. Discuss whether the following processes are good or bad ideas, and why
   - Dividing all features by the standard deviation of the training set
   - Dividing all features by the standard deviation of the test set
   - Picking the best regularization parameter based on the one that gives the best test score
   - Using as many redundant features as possible in the data matrix
   - Adding huge regularization penalty in order to ensure that the test and train performance are identical
5. **Linear regression** What is the condition number of a matrix? Is a large / small number good/bad? Discuss why knowing that condition number may be useful, and what tricks can be used to make it "better".
6. **Support vector machines** A common depiction of the SVM problem formulation is

$$\begin{array}{ll} \underset{\theta,\xi}{\text{minimize}} & \|\theta\|_2^2 + \lambda \max\{0, \xi\} \\ \text{subject to} & y_i x_i^T \theta = 1 - \xi \end{array}$$

   - Describe each term here; namely, what do the quantities $\|\theta\|_2$ and ξ_i represent?
   - What does it mean if ξ_i is positive? negative? 0?
   - What is a *support vector*?

Linear regression

1. Suppose I have m data vectors $x_1, ..., x_m \in \mathbb{R}^n$, and corresponding labels $y_i \in \mathbb{R}$ for $i = 1, ..., m$. Describe how I use linear regression to find a fit; that is, for a new data vector x, predict the corresponding label y.
2. What is the role of 2-norm squared regularization (e.g. ridge regression) from a
   - Bayesian point of view?
   - Numerical stability point of view
   - Generalization / stability point of view?
3. In this scenario, write down the normal equations, and write out in code what you would use to solve them. (Multiple answers may apply.)
4. Show how you would incorporate 2-norm squared regularization in your model and how you would solve it.
5. **Ridge regression.** What is ridge regression? How can the hyperparameter (regression weight) be tuned to prioritize accuracy vs mitigating uncertainty?

Logistic regression

Explain this sentence: "Logistic regression is about finding the maximum likelihood estimator of the probit model."

- What is a probit model?
- What does it mean to find the maximum likelihood estimator?
- What do you actually "do" in logistic regression?

Convex optimization

1. Given a function f, write down 3 ways you can check to see if this function is convex.
2. Decide if a set is convex or not convex.
3. If f is not convex, is it appropriate to use gradient descent to minimize the function? Is it appropriate for maximizing the function? Why or why not?
4. Are the following functions convex?
   - $f(x) = x^p$ for $p = 1, 2, 3, ...$
   - $f(x) = x^p$ for $p = 1, 2, 3, ...$ over the domain $x \geq 0$
   - $f(x) = \sqrt{|x|}$
   - $f(x) = \sigma(x)$ where $\sigma(x) = 1/(1 + e^{-x})$
   - $f(x) = \log(x)$
   - $f(x) = \log(\sigma(x))$ (for same definition of σ)
   - $f(x) = \mathbb{E}_y[f(x, y)]$ where f is convex over x, for fixed y
5. If I am given a function f, and told that x is such that $\nabla f(x) = 0$ and $\nabla^2 f(x)$ is positive semidefinite, is x a local minimum of f? Why or why not?
6. In general, know the conditions for which a point is a
   - stationary point
   - local minima

- global minima

7. What is the condition number of a convex function? Why is it useful? it good or bad to have a large or small condition number? What are the consequences of having a "bad" condition number?

8. Know the iteration scheme for gradient descent

9. Know the descent lemma. (You do not need to know the proof, but do know what the step sizes have to be for descent to be 1) guaranteed and 2) upper bound optimal.)

Choose your own adventure

1. You are given a dataset for predicting whether a patient has cancer. You look at the features, and you ask a trained, experienced doctor to take a look. He/She says "Wow, based on that data, even with 25 years of experience, I can't predict anything about the patient!"

   You go back to your lab and you can

   - use up as many GPU resources as possible to train a huge neural net on this data
   - buy a crystal ball
   - ask for more/better data
   - do some feature analysis, basic statistics, to look for correlations between features and results

2. You are given a dataset of 5 temperature readings in different parts of the world. Each temperature reading shows that, over the past 50 years, temperature has gone down; yet the world is crying about global warming. Reconcile this observation.

3. You are given a dataset logging all the traffic accidents by a Tesla car. You notice that of all the logs, only .01% of them ended in accident claims. You train your model, and found that it has a 99.99% accuracy in predicting whether a log resulted in an accident. Do you

   - Break open a bottle of champagne! you solved all of car accidents!
   - something else? (describe)

Background

I will not test you on this specifically, but anything in the background slides is fair game to come up in an exam. Here are some specific tasks you should feel comfortable doing

- evaluating gradients and Hessians
- evaluating expectations and variances
- basic probability laws (law of total probability, joint vs marginal vs conditional probabilities, Bayes rule)
- i.i.d.
- distributions: uniform, Gaussian
- If there is a new distribution, I may describe part of it, and ask you to compute the rest (going from pdf to cdf, evaluating expectations, variances, etc)