Section 10: Final Review

1. Kernels

Let $\phi : d \to k$ be a feature map, and define K to be the kernel matrix of ϕ.

(a) Prove that the kernel matrix is symmetric. That is, show $K_{i,j} = K_{j,i}$.

(b) Show that K is positive semi-definite
Hint: consider the matrix B where the i^{th} column of B is $\phi(x_i)$.

2. PCA

Consider the following data set, represented as three points in $\mathbb{R}^2$. Note: in this problem we will **not** demean the dataset. Perform all calculations as if the dataset were 0 mean.

$$\begin{bmatrix} 1 & 2 \\ 1.5 & 3 \\ 6 & 12 \end{bmatrix}$$

(a) What is the first principal component vector, v_1?

(b) What is the second principal component, v_2?

(c) If we use only the first principal component to compress the dataset, what will the representation of each point be?

(d) Will this representation be lossy, or perfectly preserve the data?

Answer the same questions for the following, slightly larger dataset:

$$\begin{bmatrix} 1 & 1 \\ 1.5 & 1.5 \\ -2 & 2 \\ 4 & -4 \\ 6 & -6 \\ 2 & 2 \end{bmatrix}$$

In Lecture 17, we saw the following optimization problem in the context of autoencoders:

$$\min_{f,g} \sum_{i=1}^{n} ||x_i - g(f(x_i))||_2^2$$

Suppose we know that $f(x) = Ax$ for some matrix $A \in \mathbb{R}^{n \times d}$ and $g(y) = By$ for some matrix $B \in \mathbb{R}^{d \times n}$.

(a) How could you calculate A, B?

(b) In the special case that $d = n$ what happens?

3. SVD

We will now explore SVD to transform matrices into another form. Recall that we can decompose any matrix by SVD into 3 components USV^T where matrices U and V^T are othornornal square matrices and S is a rectangular diagonal matrix.

Here we will use an example from lecture on kernels (Lecture 12). In that lecture we considered the optimization problem

$$\arg\min_w \sum_{i=1}^{n} (y_i - \phi(x_i)^T w)^2 + \lambda ||w||_2^2$$

, and we said that both of the following expressions were closed forms for the optimum $\hat{w}$:

$$\hat{w} = (\Phi^T \Phi + \lambda I_p)^{-1} \Phi^T y$$

and

$$\hat{w} = \Phi^T (\Phi \Phi^T + \lambda I_n)^{-1} y$$

Let USV^T be the SVD decomposition of Φ. Using the second of those two expressions, show $\hat{w} = VS^T(SS^T + \lambda I)^{-1}U^T y$. (This exercise is the first step in the proof that those two forms are equivalent – to complete the proof you would show the other expression can also be rewritten in this form).

4. More SVD, Gaussians

Let Σ be a 2×2 matrix with eigenvalues $v_1 = [1/2, \sqrt{3}/2]^T$ with eigenvalue 3 and $v_2 = [-\sqrt{3}/2, 1/2]^T$ with eigenvalue 2.

(a) Give an expression for Σ (Hint: look at the title of this section)

(b) Let $\Sigma^{1/2}$ be the matrix such that $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. Give an expression for $\Sigma^{1/2}$. Do not try to simplify your expression.

(c) Recall that the density of a multivariate Gaussian with mean μ and covariance Σ is:

$$P(x) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

Where $|.|$ is the determinant operator. Let $\mathcal{E}$ be the set of all points where the density is $e^{-3/2}/(2\sqrt{6}\pi)$, i.e. $\mathcal{E} = \{x | P(x) = e^{-3/2}/(2\sqrt{6}\pi)\}$ ($\mathcal{E}$ is an "isocontour"). Derive a simple formula for $\mathcal{E}$ (where μ is unknown, but Σ is the matrix from the first part). Leave your answer in terms of μ and Σ.

(d) Consider the distribution $\mathcal{N}(\mu, \Sigma)$ with $\mu = [0, 1]^T$ and Σ being the matrix from part a. Draw the contour you described in the previous part. Hint: you should only need to think about the eigenvectors and eigenvalues of Σ to figure out what to draw.

The numbers defining $\mathcal{E}$ are not nice enough to graph by hand; it's enough to figure out its shape and the formulas for enough important points to make a plot.

Neither of these matrices can the be the covariance of a Gaussian. For each, give a reason why it's not.

(a) $\begin{bmatrix} 3 & 4 \\ 4 & -1 \end{bmatrix}$

(b) $\begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}$

In section we proved directly that for any matrix A, AA^T is positive semi-definite. Use the SVD of A to show the same result.

5. EM

Suppose that Kevin and Anna assign grades as follows: With probability 0.5 you get a 4.0, with probability μ you get a 3.7, with probability 2μ you get a 3.4 and with probability $0.5 - 3\mu$ you get a 3.0. (μ is assumed to be between 0 and $1/6$.)

(a) You wish to find the maximum likelihood estimate of μ from data. You somehow managed to get ahold of some data and found out that a 4.0's were assigned, b 3.7's were assigned, c 3.4's were assigned and d 3.0's were assigned. What is the MLE of μ given a, b, c, d?

(b) Now suppose that instead some information is hidden. Specifically, you are told c and d (the number of 3.4's and 3.0's), but you only know h which is the combined number of 4.0's and 3.7s. Describe how you would use EM to solve for μ by filling in the following:

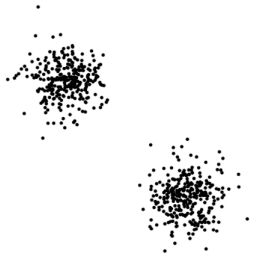
E step: if you knew the value of μ, you could compute the expected value of a and b.

M step: If you knew the expected values of a and b, you could compute the MLE of μ.

6. Clustering Methods

You want to cluster your data into 2 clusters. For each of the following datasets, which of k-means and EM for Gaussian Mixture Models would provide a faithful representation? Answer k-means, GMM, Both, Neither.

(a)



(b)

7. Linear Regression

7.1. Regularization

You're running linear regression to predict from one-dimensional feature vectors. Your training set consists of the points $(2, 2), (4, 4), (2, 1)$ (where the first entry is the feature, and the second entry is the value to be predicted).

You are going to learn w and b, and use the predictor $y = wx + b$ (note that we don't need a transpose on w, because it is actually a single number, not a vector like usual).

You are deciding on a regularization strategy. You try each of the following regularizers:

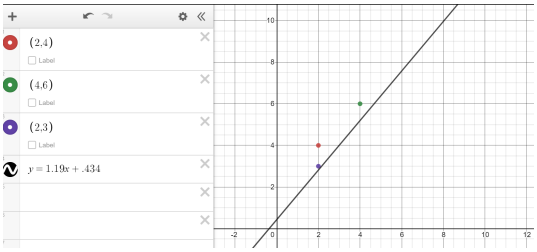
(a) no regularization

(b) $w^2 + b^2$

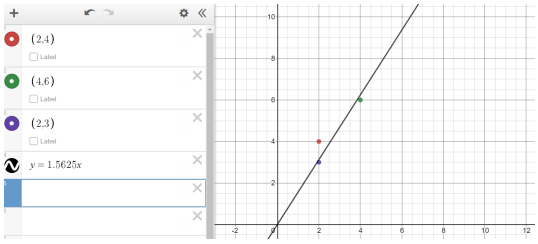
(c) $5(w^2 + b^2)$

(d) $|w| + |b|$

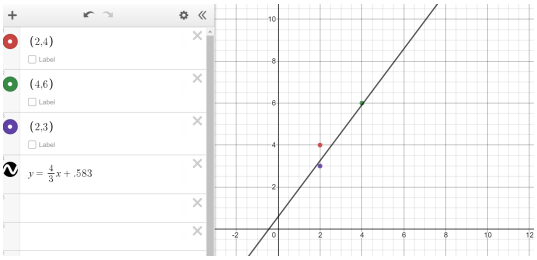
You solve all four regression problems and get the following set of predictors:



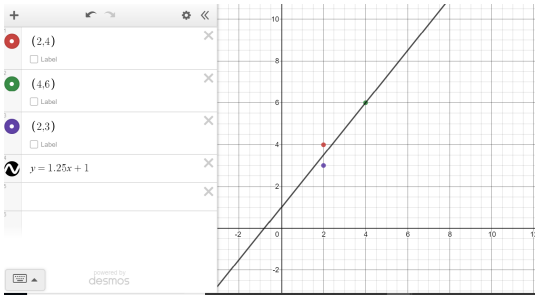
(I) $w = 1.190, b = 0.434$



(II) $w = 1.5625, b = 0.000$



(III) $w = 1.333, b = 0.583$



(IV) $w = 1.250, b = 1.000$

Answer the following questions:

(a) Match each of your regularization strategies to the predictor it created.

(b) Usually we do not regularize by b. Explain why this is usually not a good idea.

7.2. Midterm Redux: Closed Form

We will revisit the midterm problem of deriving a closed form for least squares when the errors for each sample have different (known) variances. Recall from the midterm that we are trying to minimize the negative log likelihood function of

$$\sum_{i=1}^{n} \frac{(y_i - x_i^T w)^2}{2\sigma_i^2}$$

where $x_i, w \in \mathbb{R}^{d \times 1}$ and $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{Y} \in \mathbb{R}^{n \times 1}$

Write the gradient of this function with respect to w.

Find a closed form of the best w (you will need to convert to matrix form to do this easily).

7.3. Alternative Noise Model

Recall that if we assume Gaussian noise on the labels then the MLE of our estimator w is one that minimizes the residual sum of squares $\sum_{i=1}^{n}(y_i - x_i^T w)^2$.
Instead, let's model the noise as a Laplace distribution. The PDF of Laplace(μ, b) is

$$P(y|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|y - \mu|}{b}\right)$$

(a) Write the likelihood function for the dataset given training data $\{(x_i, y_i)\}_{i=1}^n$

(b) Write the negative log-likelihood and show the MLE is the w that minimizes the *sum of absolute residuals*

$$\sum_{i=1}^{n} |y_i - x_i^T w|$$

(c) Write the gradient descent update rule for minimizing the sum of absolute residuals.

(d) Why might we want to use this noise model instead of Gaussian noise?

8. SVMs

This problem relies on a very specific definition of "support vector" that we have not discussed in class, so you do not need to worry about it.

Consider the dataset consisting of 7 data points, 4 with positive labels $\{0, 1, 2, 3\}$, and 3 with negative labels $\{-3, -2, -1\}$. Suppose we want to learn a linear SVM with slack variables for this dataset. Recall we can formalize this as a constrained optimization problem:

$$\min_{w,b,\xi} ||w||^2 + C \sum_i \xi_i$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \; \forall i$$

$$\xi_i \geq 0 \; \forall i$$

where C is a regularization parameter that balances the size of the margin (smaller $||w||^2$) vs. the violation of the margin (smaller $\sum_i \xi_i$).

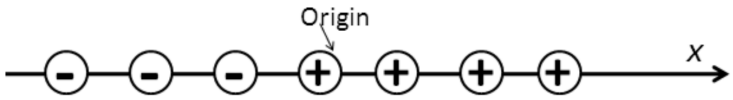


Figure 1: Dataset

~~Also recall that a support vector is a datapoint that lies on a margin.~~

(a) If $C = 0$, which means we only care about the size of the margin, how many support vectors do we have?

(b) If $C \to \infty$, which means we only care about the violation of the margin, how many support vectors do we have?

9. Optimization Basics

(a) Why do we use the word "convex" to refer to both convex sets and convex functions (i.e. why are they related?)

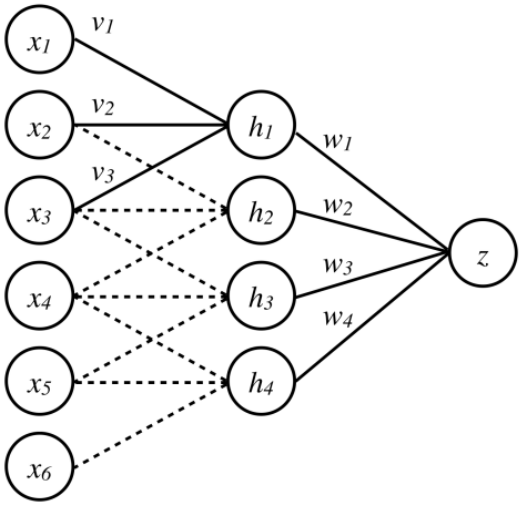
(b) We've talked about convex sets of points in $\mathbb{R}^n$, but we can talk about convex subsets of other spaces. Instead of considering subsets of $\mathbb{R}^n$, let's consider subsets of $\mathbb{R}^{n \times n}$. I.e. instead of sets of points, we now have sets of matrices. In this ambient space, is the set of symmetric matrices a convex set?

(c) For each of the following, either give a one-sentence explanation for why the statement is true, or a counter-example.

- If f and g are convex functions, then $f + g$ is a convex function.
- If $f + g$ is a convex function, then f and g are both convex functions.
- If A is a convex set then $A \cap B$ is convex for any set B.
- If A and B are convex and $A \cap B \neq \varnothing$ then $A \cup B$ is a convex set.

10. Neural Networks

Consider the convolutional neural network architecture:



In the first layer, we have a one-dimensional convolution with a single filter of size 3 such that $h_i = \sigma(\sum_{j=1}^{3} v_j x_{i+j-1})$. The second layer is fully connected, such that $z = \sum_{i=1}^{4} w_i h_i$. The hidden units' activation function $\sigma(x)$ is the sigmoid function. The output unit is linear. We perform gradient descent on the loss function $L = (y - z)^2$ where y is the training label for x.

(a) How many parameters does this network have? Recall that convolutional layers share weights. There are no bias terms.

(b) Why might we use convolutions?

Consider the neural network defined as $g(x) = W_2 f(W_1 x)$. Where f is an activation function, $W_1 \in \mathbb{R}^{h \times d}$, $W_2 \in \mathbb{R}^{1 \times h}$. Let us consider minimizing the square loss for a regression task.

(c) Let $f(x) = \tanh(x)$. Is the loss function convex in the model parameters?

(d) Let $f(x) = x$. What types of functions can we represent using this model?

11. Multiple Choice T/F and Short Answer

(a) Let X have the following Singular Value Decomposition: $X = USV^T$. For each of the follow matrices, state whether or not the columns of U are the eigenvectors of that matrix.

- $X^T X$
- $X^T X X^T X$
- $X X^T$
- $X X^T X X^T$

(b) Which of the following are reasons why PCA useful for doing preprocessing steps on data? Please provide justification
a) Reduce overfitting of data
b) Increase computational efficiency during training
c) Extract more features/information about your training data

(c) Let us consider a matrix $X = \sum_{i=1}^{r} s_i u_i v_i^T$. What are the possible values for the rank for this matrix?

(d) Which of the following are true statements for k-nearest neighbors as we increase k?
a) The decision boundary becomes smoother
b) Variance decreases
c) As the number of sample points approaches infinity, error rate becomes less than twice the Bayes error for 1-nearest neighbors.
d) Bias decreases

(e) In lecture we have shown that low-rank approximation of matrix is helpful for many reasons. Which of the following are the reason to use a low-rank appoximation of your data matrix?
a) Denoising
b) Discover latent categories in data
c) Filling in unknown values
d) Storage Compression

(f) What is the leave one out cross-validation error when using an SVM for classification on the following example? (In this case, error will simply be the number of mistakes.) The SVM decision boundary and margin on either side are shown.

(g) What is the minimal number of points we can remove required to change the decision boundary?

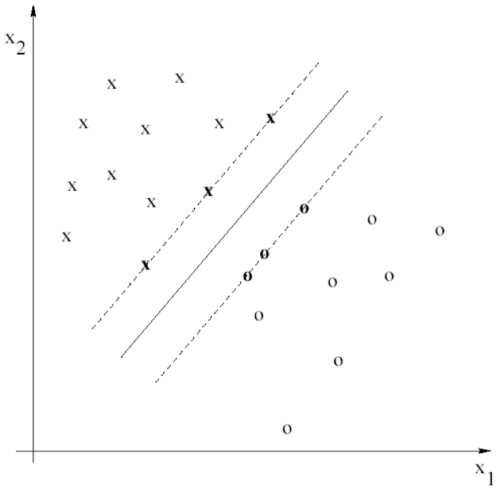
(h) There are several ways to formulate the hard margin SVM. Consider a formulation where we try to directly maximize the margin γ. The training samples are $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and their labels are $y_1, \ldots, y_n$. Circle all the constraints we should impose to get a correct SVM. Maximize γ subject to

(i) $y_i \mathbf{x}_i^T \mathbf{w} \leq \gamma \quad \forall i$

(ii) $y_i \mathbf{x}_i^T \mathbf{w} \geq \gamma \quad \forall i$

(iii) $||\mathbf{w}|| \geq 1$

(iv) $||\mathbf{w}|| = 1$



12. Matchup

Here's a sample of some of the techniques we've seen in this course:

(a) Bootstrap

(b) EM Algorithm

(c) PCA/SVD

(d) Kernelization

(e) cross-validation

For each of the following scenarios, choose the most appropriate of the above techniques.

(a) You have a large number of features, and want to find the most important linear combinations of your features.

(b) You think your linear model would work best if you applied a feature map, but it would be infeasible to calculate all the entries of the map.

(c) You're worried your model has too much variance, and want to estimate how much your predictions would change if you got a different training set.

(d) Your model uses α and β. The best α depends on the best β and the best β depends on the best α. You can't find a closed form that lets you find both α and β simultaneously, but given some $\hat{\beta}$ you can find the best $\hat{\alpha}$ and vice-versa. How do we find α, β?

(e) Your model uses α and β. You aren't interested in finding the best α for the training set, but in finding one that will allow for better generalization. How do we find α?

13. Midterm Matchup Redux

This problem is copy-pasted from the midterm. The solutions will have a detailed breakdown of this question; you may want to try it again to see what you remember from the first half of the course.

Suppose you are doing regression and discover that your learning algorithm is not working as well as you think it should be. To fix the problem, some things you could try are the following:

(a) Get a hold of a larger training set.

(b) Add additional functions to your hypothesis class (e.g., switch from linear functions to cubic polynomials.)

(c) Try using a simpler hypothesis class

(d) Change the feature map

(e) Try changing the optimization algorithm used.

(f) Try L2 regularization

(g) Try L1 regularization

For each of the following specific issues, which of the above approaches has the potential to help?

For each of these problems, if there are two or more strategies that will help, list the two that you think are most likely to help. If there is only one strategy that will help, list it. If there are none, write "none".

(i) high bias

(ii) high variance

(iii) high irreducible noise

(iv) overfitting

(v) underfitting

(vi) difficulty understanding which features are most important for prediction.