MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering and Computer Science
6.036—Introduction to Machine Learning
Spring Semester 2017

Assignment 3

Issued: Tuesday, February 28th
Due: Friday, March 10th, 9:00 AM

Collaborative Filtering, Kernels, Linear Regression

1. **Collaborative Filtering**

   In this question, we will use the alternating projections algorithm for low-rank matrix factorization, which aims to minimize

   $$J(U,V) = \frac{1}{2}\sum_{(a,i)\in D}(Y_{ai} - [UV^T]_{ai})^2 + \frac{\lambda}{2}\sum_{a=1}^{n}\sum_{j=1}^{k}U_{aj}^2 + \frac{\lambda}{2}\sum_{i=1}^{m}\sum_{j=1}^{k}V_{ij}^2$$

   Let Y be defined as

   $$Y = \begin{bmatrix} 5 & ? & 7 \\ ? & 2 & ? \\ 4 & ? & ? \\ ? & 3 & 6 \end{bmatrix}$$

   and let $k = \lambda = 1$. Additionally, U and V are initialized as $U^{(0)} = [6, 0, 3, 6]^T$, and $V^{(0)} = [4, 2, 1]^T$.

   (a) Compute X, the matrix of predicted rankings.

   (b) Compute the squared error and the regularization terms for the current estimate X.

   (c) Suppose V is kept fixed. Run one step of the algorithm to find the new estimate $U^{(1)}$.

   (d) What would happen to $U^{(1)}$ if λ were increased from 1 to a very large value?

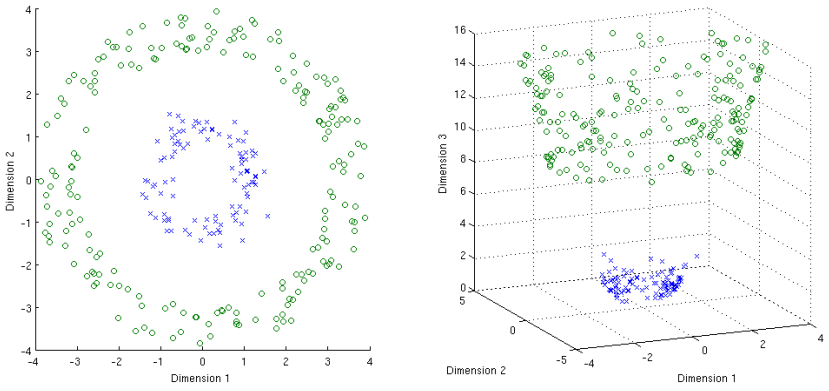
2. **Kernels**

(a) Let $x, q \in \mathbb{R}^2$ be two feature vectors, and let $K(x, q) = (x^T q + 1)^2$. This is often known as a polynomial kernel. It's simple to compute: you just take the dot product between two feature vectors, add one, and then square the result. But what kind of feature mapping does this kernel implicitly use? Assuming we can write $K(x, q) = \phi(x)^T \phi(q)$, derive an expression for $\phi(x)$. As a simple example that uses this kernel, imagine that our feature vectors were bag of words vectors. In this example, give an intuitive interpretation of what the $\sqrt{2}x_1x_2$ term in the expression for $\phi(x)$ you just wrote down means.

(b) In the figure below, a set of points in 2-D is shown on the left. On the right, the same points are shown mapped to a 3-D space via some transform $\phi(x)$, where x denotes a point in the 2-D space. Notice that $\phi(x)_1 = x_1$ and $\phi(x)_2 = x_2$, or in other words, the first and second coordinates are unchanged by the transformation.

i. Which of the following functions could have been used to compute the value of the 3rd coordinate, $\phi(x)_3$ for each point: $\phi(x)_3 = x_1 + x_2$, $\phi(x)_3 = x_1^2 + x_2^2$, $\phi(x)_3 = x_1x_2$, or $\phi(x)_3 = x_1^2 - x_2^2$?

ii. Think about how a linear decision boundary in the 3 dimensional space ($\{\phi \in \mathbb{R}^3 : \theta \cdot \phi + \theta_0 = 0\}$) might appear in the original 2 dimensional space. Approximately draw the resulting decision boundary (not necessarily linear) in the original 2-D space. In other words, plot $\{x \in \mathbb{R}^2 : \theta \cdot \phi(x) + \theta_0 = 0\}$).



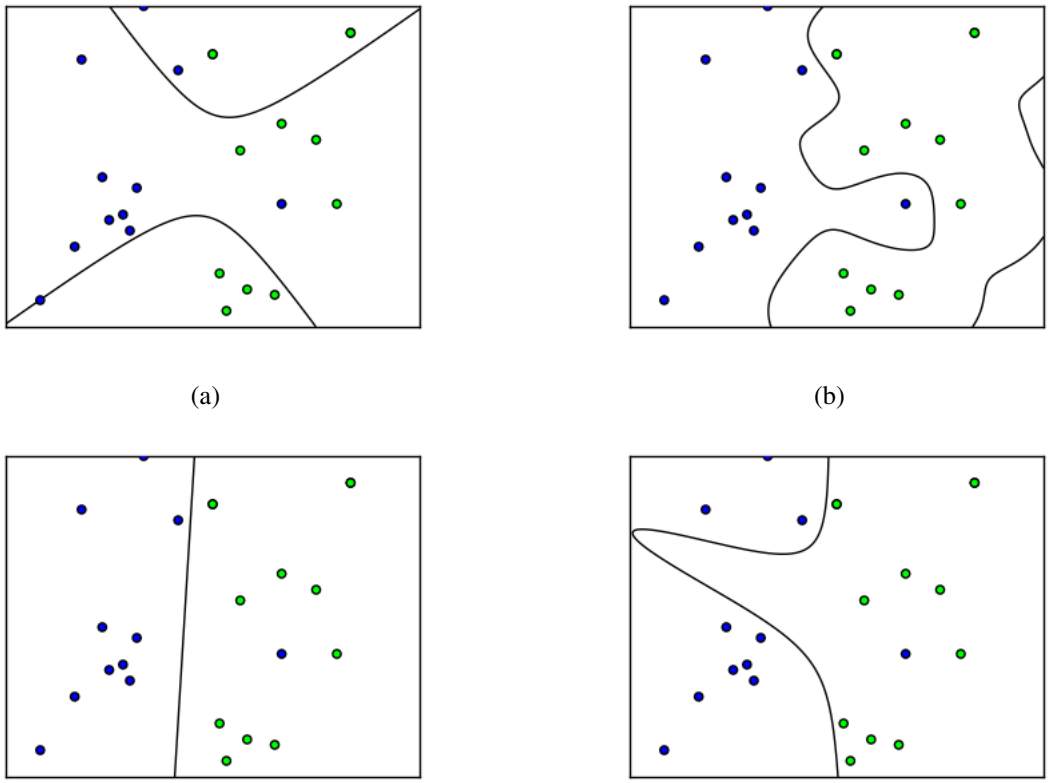
(c) Consider fitting a kernelized SVM to a dataset $(x^{(i)}, y^{(i)})$ with $\forall i \; x^{(i)}, y^{(i)} \in \mathbb{R}$. To fit the parameters of this model, one computes θ and θ_0 to minimize the following objective:

$$L(x; \theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \text{Loss}_h(y^{(i)}(\theta \cdot \phi(x^{(i)}))) + \frac{\lambda}{2}||\theta||^2$$

where ϕ is the feature vector associated with the kernel function. Note that, in a kernel method, the optimization problem for training would be typically expressed solely in terms of the kernel function $K(x, x')$ (dual) rather than using the associated feature vectors $\phi(x)$ (primal). We use the primal only to highlight the classification problem solved.

The plots in Figure 1 show 4 different kernelized SVM models estimated from the same 11 data points. We used a different kernel to obtain each plot but got confused about which plot corresponds to which kernel. Help us out by assigning each plot (a-d) to one of the following models (i-iv). Explain your reasoning and describe qualitatively how the resulting classifiers vary with the value of λ.

i. Linear kernel
ii. Quadratic kernel
iii. 3rd-order kernel
iv. Gaussian RBF kernel



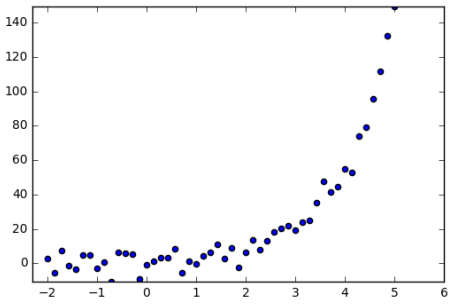
(a) (b)

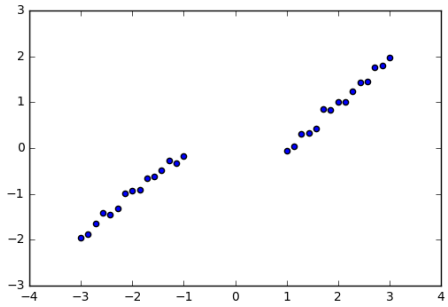
(c) (d)

Figure 1: Classifiers found by SVM for different choices of kernel

3. **Linear Regression and Regularization**

(a) For each of the datasets below, provide a simple feature mapping ϕ such that the transformed data $(\phi(x^{(i)}), y^{(i)})$ would be well modeled by linear regression.

i. 

ii. 

(b) Consider fitting a ℓ_2-regularized linear regression model to data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ where $x^{(t)}, y^{(t)} \in \mathbb{R}$ are scalar values for each $t = 1, \ldots, n$. To fit the parameters of this model, one solves $\min_{\theta \in \mathbb{R},\, \theta_0 \in \mathbb{R}} \{L(\theta, \theta_0)\}$ where

$$L(\theta, \theta_0) = \sum_{t=1}^{n} (y^{(t)} - \theta x^{(t)} - \theta_0)^2 \;\; + \;\; \lambda \theta^2$$

Here $\lambda \geq 0$ is a pre-specified fixed constant, so your solutions below should be expressed as functions of λ and the data. This model is typically referred to as *ridge regression*.

i Write down an expression for the gradient of the above objective function in terms of θ, θ_0.

ii Find the closed form expression for θ_0 and θ which solves the ridge regression minimization above.