Trial Midterm

Introductiont to Machine Learning
Fall 2018
Instructor: Anna Choromanska

Problem 1

Assume we are given N pairs of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, where each x and y is just a scalar and we wish to do a simple 1-dimensional linear regression with the function below:

$$f(x) = \theta_0 + \theta_1 x.$$

Assume we have the means of both the x and y, i.e.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i; \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

We find the optimal setting of θ_0^* and θ_1^* by minimizing the squared error:

$$L(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \theta_0 - \theta_1 x_i)^2.$$

a) Put an "X" besides each statement that is true when we have the optimal least squared error setting for our parameters θ_0^* and θ_1^*:

( ) $\frac{1}{N}\sum_{i=1}^{N}(y_i - \theta_0^* - \theta_1^* x_i)y_i = 0$
( ) $\frac{1}{N}\sum_{i=1}^{N}(y_i - \theta_0^* - \theta_1^* x_i)(y_i - \bar{y}) = 0$
( ) $\frac{1}{N}\sum_{i=1}^{N}(y_i - \theta_0^* - \theta_1^* x_i)(x_i - \bar{x}) = 0$
( ) $\frac{1}{N}\sum_{i=1}^{N}(y_i - \theta_0^* - \theta_1^* x_i)(\theta_0^* + \theta_1^* x_i) = 0$

b) Suppose we have the following components of the Gaussian sufficient statistics from the data. Show how we could compute the optimal value of θ_1^* only by using two of the following 5 scalar numbers:

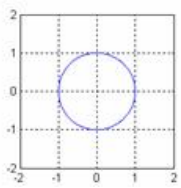
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i; \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$C_{xx} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2; \quad C_{yy} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2; \quad C_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}).$$

Problem 2: VC dimension

We are forming a two-dimensional classifier in the space.

a) Consider the following circular classifier in $2d$: $f(x) = \text{sign}(x^\top x + b) = \text{sign}(x_1^2 + x_2^2 + b)$, where $x = [x_1, x_2]^\top$. Here, the parameter of the classifier is b, just a scalar. What is the VC dimension h of



this quadratic classifier in two dimensional classification problems? Justify your answer with appropriate drawings. Next, draw a poor configuration for h points that the quadratic classifier cannot shatter and show the particular labelings of the points that it cannot shatter ("shatter" means that the quadratic classifier will perfectly separate the data despite an arbitrary binary labeling of the points).

b) Redo the tasks from a) when the classifier under consideration instead has the form: $f(x) = \text{sign}(ax^\top x + b) = \text{sign}(ax_1^2 + ax_2^2 + b)$. Now, the parameters of the classifier are a and b, which are both just scalars. HINT: think about what happens when a goes negative!

Problem 3

Consider the following plot, where we fit the polynomial of order M ($f(x; w) = \sum_{j=0}^{M} w_j x^j$) to the dataset, where $w = [w_0 \quad w_1 \quad \dots \quad w_M]^\top$ denotes the vector of model weights and the dataset is a collection of 2-dimensional points (x, y). The dataset is represented with the blue circles on the figure.

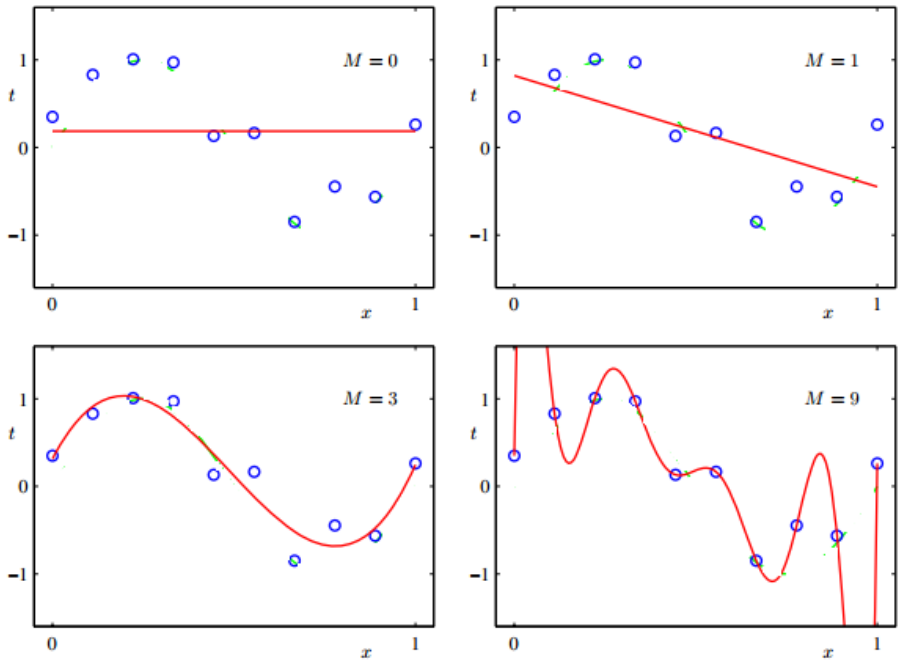


Figure 1: Plots of polynomials having various orders M, shown as red curves, fitted to the data set.

What is the reasonable choice of M and why? Which M correspond to overfitting and which to underfitting and why?

Consider any loss function that measures the discrepancy between the target values and the predictions of the model, e.g. squared loss which for a single data point is defined as $L(y_i, f(x_i, w)) = \frac{1}{2}(y_i - f(x_i, w))^2$). Draw a typical behavior of the train and test loss for the optimal setting of model weights as a function of M, where recall that the train loss is the loss computed for a training dataset (the model was trained on this dataset) and the test loss is the loss computed for a test dataset (the model did not see this dataset during training). Indicate overfitting and underfitting regimes.

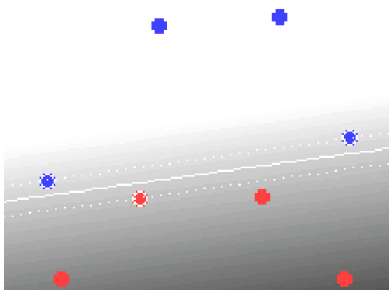
Problem 4

Consider $x_1, x_2, \ldots, x_N$ to be N observed values of the i.i.d. random variables $X_1, X_2, \ldots, X_N$ from Poisson distribution, i.e. $P(x) = e^{-\lambda}\frac{\lambda^x}{x!}$. Compute maximum likelihood estimate for λ.

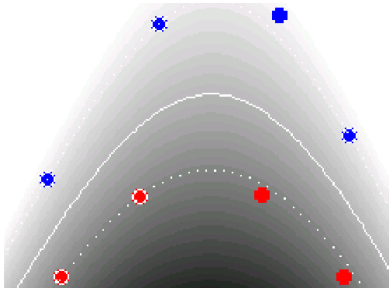
Problem 5

Assume we have trained 3 separable support vector machines on the 2D data (the axes go from -1 to 1 in both horizontal and vertical direction) using 3 different kernels as follows:

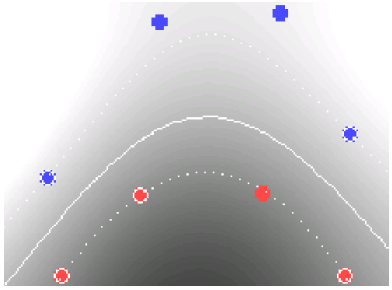
a) a linear kernel (i.e. the standard linear SVM): $k(x_i, x_j) = x_i^\top x_j$



b) a quadratic polynomial kernel: $k(x_i, x_j) = (x_i^\top x_j + 1)^2$



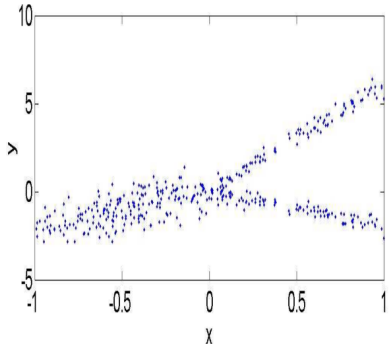
c) an RBF kernel: $k(x_i, x_j) = \exp\left(-\frac{1}{2\sigma}\|x_i - x_j\|^2\right)$



Assume we now translate the data by adding a large constant value (i.e. 10) to the vertical coordinate of each data points, i.e. a point (x_1, x_2) becomes $(x_1, x_2 + 10)$. If we retrain the above SVMs on this new data, how does the resulting SVM boundary change relative to the data points? Explain why or why not it changes for all 3 cases (a), (b), and (c) and draw what happens to the resulting new boundaries when appropriate.

Problem 6

You are given the data set in the figure below which is fit with maximum likelihood via EM using a mixture of 3 Gaussians. Assume EM converged nicely to the optimal maximum likelihood solution. Draw the 3-Gaussian fit you would expect on top of the data below. (10 points)



EM thus gives us the following joint distribution:

$$p(x,y) = \sum_m p(m,x,y) = \sum_{m=1}^{3} \alpha_m \mathcal{N}(x,y|\mu_m,\Sigma_m).$$

The mixture model is conditioned to form a mixture of experts conditional pdf:

$$p(y|x) = \sum_m p(m,y|x) = \sum_{m=1}^{3} p(m|x)p(y|m,x).$$

Assume the original Gaussians give rise to Gates $p(m|x)$ functions as above and the conditioned Gaussians give rise to the Experts $p(y|m,x)$. In the 3 figures below, draw three expert/gate combinations (i.e. $p(y|x,m)$ and $p(m|x)$) for $m=1$, $m=2$, and $m=3$. The order ($m=1,2,3$) of the experts/gates doesn't matter. Plot each expert as a contour plot of the conditional probability of y given m and x as x,y varies and plot the value of $p(m|x)$ for each gate as x varies. (30 points) Briefly explain your answer. (10 points)

$p(y|x,m=1)$ & $p(m=1|x)$ $p(y|x,m=2)$ & $p(m=2|x)$ $p(y|x,m=3)$ & $p(m=3|x)$