CPSC340: Additional Exercises

Exercise 1. The following short questions should be answered with at most two sentences, and/or a picture. For the (true/false) questions, answer true or false. If you answer true, provide a short justification, if false explain why or provide a small counterexample.

- In one sentence, characterize the differences between classification and regression.

Answer: Classification maps inputs to discrete outputs whereas regression maps inputs to continuous outputs.

- Your billionaire friend needs your help. She needs to classify job applications into good/bad categories, and also to detect job applicants who lie in their applications using density estimation to detect outliers. To meet these needs, do you recommend using a discriminative (e.g. logistic) or generative classifier? Why?

Answer: If you want to use density estimation to detect outliers in the applications (e.g. input x^i) then you need probabilistic models on the input, so you need a generative classifier.

- Your billionaire friend also wants to classify software applications to detect bug-prone applications using features of the source code. This pilot project only has a few applications to be used as training data, though. To create the most accurate classifier, do you recommend using a discriminative or generative classifier? Why?

Answer: Discriminative as there are too few data to learn reliably the class conditional densities of the training data.

- Finally, your billionaire friend also wants to classify companies to decide which one to acquire. This project has lots of training data based on several decades of research. To create the most accurate classifier, do you recommend using a discriminative or generative classifier? Why?

Answer: Generative as a lot of traning data are available so we can expect to learn properly the class conditional densities of the training data.

- Assume that we are using some classifier of fixed complexity. How will the test error and cross-validation error typically behave as the number of training examples increase?

Answer: They will be typically decreasing and then stabilize.

- Both PCA and linear regression can be thought of as algorithms for minimizing a sum of squared errors. Explain which error is being minimized in each algorithm.

Answer: PCA minimizes

$$\sum_{i=1}^{N} \|\mathbf{x}_i - \widehat{\mathbf{x}}_i\|^2$$

where $\widehat{\mathbf{x}}_i = \sum_{j=1}^{k} \left(\mathbf{x}_i^{\mathrm{T}} \mathbf{u}_j\right) \mathbf{u}_j$ whereas linear regression minimizes

$$\sum_{i=1}^{N} \left\|y_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right\|^2$$

- Consider a real-valued random variable X admitting a continuous probability density function $f(x)$. Is the probability that $X = x$ equal to $f(x)$?

Answer: False. The probability that $X = x$ is equal to zero.

- Besides EM, is it possible to use gradient descent to perform inference or learning on a Gaussian mixture model?

Answer: It is entirely possible to use gradient descent.

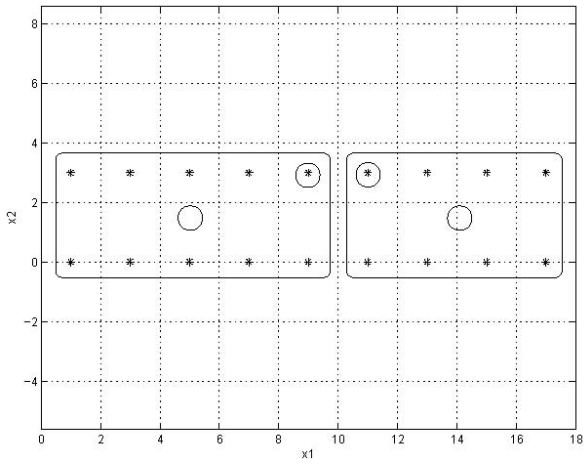


Figure 1: K-means

Exercise 2. Consider the data plotted in Figure 2a, which consist of two rows of equally spaced points. If K-means clustering (K =2) is initialised with the two points whose coordinates are (9, 3) and (11, 3), indicate the final clusters obtained (after the algorithm converges).

Answer is given in the graph.

Exercise 3. Consider a regression problem where the two dimensional input points $\mathbf{x} = (x_1\ x_2)^{\mathrm{T}}$ are constrained to lie within the unit square: $x_i \in [-1, 1]$, $i = 1, 2$. The training and test input points $\mathbf{x}$ are sampled uniformly at random within the unit square. The target outputs y are governed by the following model

$$y \sim \mathcal{N}\left(x_1^3 x_2^5 - 10x_1x_2 + 7x_1^2 + 5x_2 - 3, 1\right).$$

In other words, the outputs are normally distributed with mean given by

$$x_1^3 x_2^5 - 10x_1x_2 + 7x_1^2 + 5x_2 - 3$$

and variance 1.

We learn to predict y given $\mathbf{x}$ using linear regression models with 1st through 10th order polynomial features. The models are nested in the sense that the higher order models will include all the lower order features. The estimation criterion is the mean squared error. We first train a 1st, 2nd, 8th, and 10th order model using $N = 20$ training points, and then test the predictions on a large number of independently sampled points.

Select all the appropriate model(s) for each column. If you think the highest, or lowest, error would be shared among several models, be sure to list all models.

	Lowest Training error	Highest Training error	Lower test error (typically)
1st order			
2nd order			
8th order			
10th order			

Answer. We have

	Lowest Training error	Highest Training error	Lower test error (typically)
1st order		X	
2nd order			X
8th order	X		
10th order	X		

The 10th order regression model would seriously overfit when presented only with $N = 20$ training points. The second order model on the other hand might find some useful structure in the data based only on 20 points. The true model is also dominated by the second order terms. Since $x_i \in [-1, 1]$ for $i = 1, 2$ any higher order terms without large coefficients are vanishingly small.

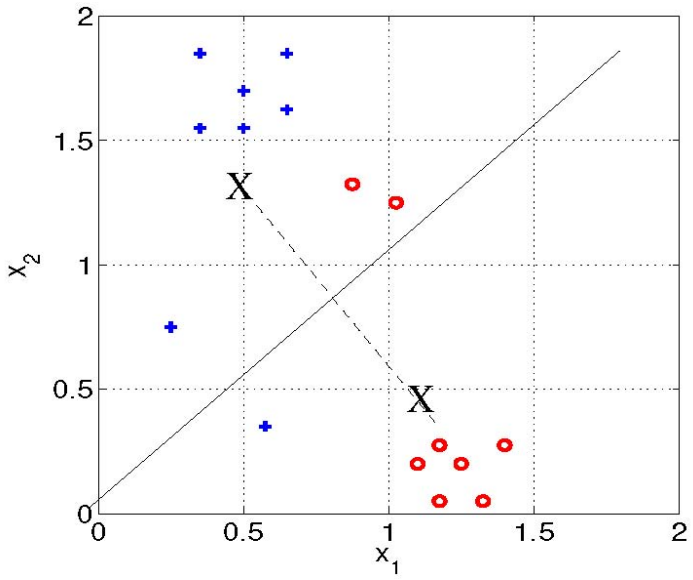


Figure 2: Labeled training set, where blue crosses resp. red circles corresponds to class $y = 1$ resp. $y = 0$

Exercise 4. We consider here generative and discriminative approaches for solving the classification problem illustrated in the following figure. Specifically, we will use a mixture of Gaussians model and regularized logistic regression models.

1. We will first estimate a mixture of Gaussians model, one Gaussian per class, with the constraint that the covariance matrices are identity matrices. The mixing proportions (class frequencies) and the means of the two Gaussians are free parameters.

- Sketch the maximum likelihood estimates of the means of the two class conditional Gaussians. Mark the means as points "x" and label them "0" and "1" according to the class.

Answer: See graph, the means should be close to the center of mass of the points.

- Draw the decision boundary in the same figure.

Answer: See graph. Since the two classes have the same number of points and the same covariance matrices, the decision boundary is a line and, moreover, should be drawn as the orthogonal bisector of the line segment connecting the class means (see lecture 18, slide 15).

2 We have also trained regularized linear logistic regression models

$$P(y = 1|\mathbf{x}, \mathbf{w}) = g(w_0 + w_1 x_1 + w_2 x_2)$$

for the same data; i.e. g is the logistic function. The regularization penalties, used in penalized conditional log-likelihood estimation, were $-Cw_i^2$, where

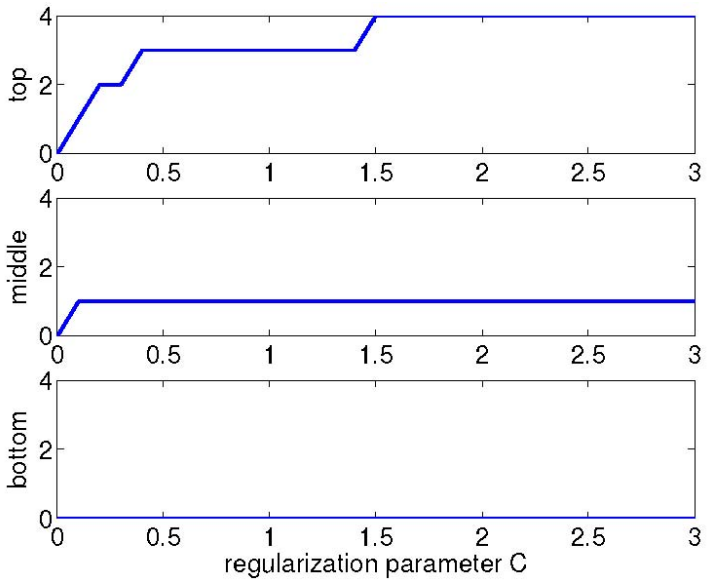


Figure 3: Training errors as a function of regularization penalty

$i = 0, 1, 2$. In other words, only one of the parameters were regularized in each case. Based on the classification data, we generated three plots, one for each regularized parameter, of the number of misclassified training points as a function of C (Figure 4.2). The three plots are not identified with the corresponding parameters, however. Please assign the "top", "middle", and "bottom" plots to the correct parameter, w_0, w_1, or w_2, the parameter that was regularized in the plot. Provide a brief justification for each assignment.

- What is the "top"?

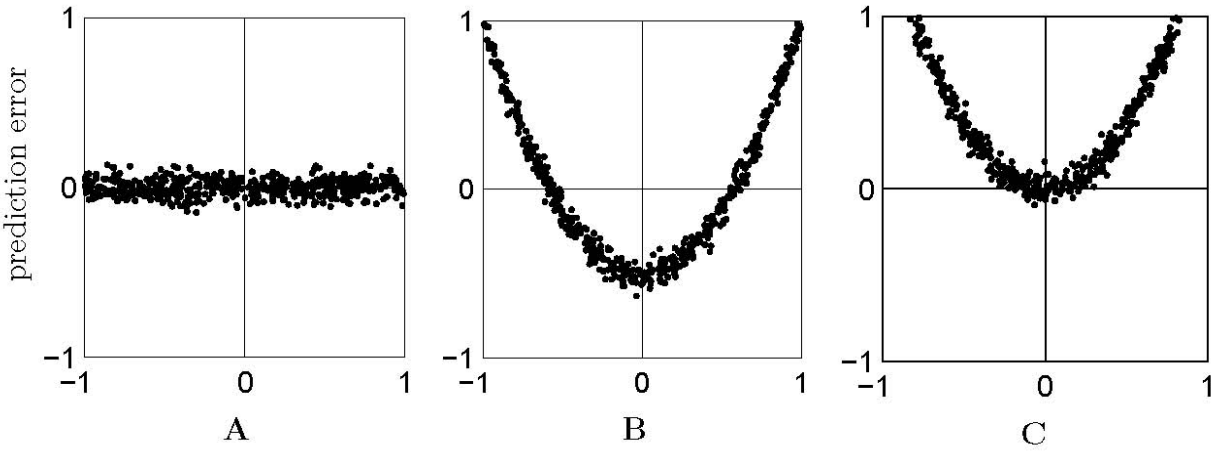
Answer: "top"= w_1. By strongly regularizing w_1 we force the boundary to be horizontal in the figure. The logistic regression model tries to maximize the log-probability of classifying the data correctly. The highest penalty comes from the misclassified points and thus the boundary will tend to balance the (worst) errors. In the figure, this is roughly speaking $x_2 = 1$ line, resulting in 4 errors.

- What is the "middle"?

Answer: "middle"= w_0. If we regularize w_0, then the boundary will eventually go through the origin (bias term set to zero). Based on the figure we can find a good linear boundary through the origin with only one error.

- What is the "bottom"?

Answer: "bottom"= w_2. The training error is unaffected if we regularize w_2 (constrain the boundary to be vertical); the value of w_2 would be small already without regularization.

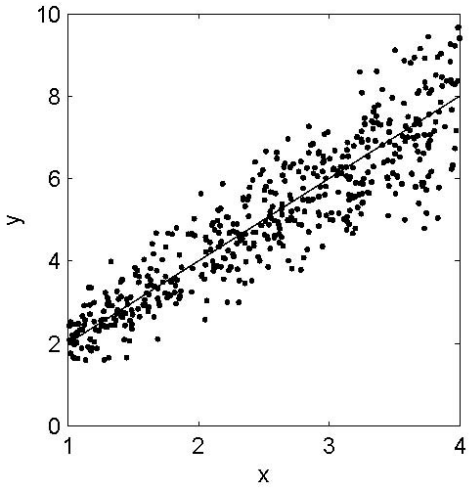


Exercise 5. Each plot above claims to represent prediction errors as a function of x for a trained regression model based on some dataset. Some of these plots could potentially be prediction errors for linear or quadratic regression models, while others couldn't. The regression models are trained with the least squares estimation criterion. Please indicate compatible models and plots.

	A	B	C
linear regression			
quadratic regression			

Answer.

	A	B	C
linear regression	x	x	
quadratic regression	x		



Exercise 6. Here we explore a regression model where the noise variance is a function of the input (variance increases as a function of input). Specifically

$$y = wx + \varepsilon$$

where the noise is normally distributed with mean 0 and standard deviation σx. The value of σ is assumed known and the input x is restricted to the interval $[1, 4]$. We can write the model more compactly as

$$y \sim \mathcal{N}(wx, \sigma^2 x^2).$$

If we let x vary within $[1, 4]$ and sample outputs y from this model with some w, the regression plot might look like the data displayed on this page.

Some potentially useful relations: if $z \sim \mathcal{N}(\mu, \sigma^2)$, then $az \sim \mathcal{N}(a\mu, a^2\sigma^2)$ for a fixed a. If $z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and are independent then $Var\,(z_1 + z_2) = \sigma_1^2 + \sigma_2^2$.

- What is the probability density function of the ratio y/x for a fixed value of x?

Answer: Since we have $y \sim \mathcal{N}(wx, \sigma^2 x^2)$ then

$$y/x \sim \mathcal{N}(w, \sigma^2).$$

Suppose we now have N independent training data $\left\{\left(x^i, y^i\right)\right\}_{i=1}^N$ where each x^i is chosen at random from $[1, 4]$ and the corresponding y^i is subsequently sampled from $y^i \sim \mathcal{N}(w^* x^i, \sigma^2 \left(x^i\right)^2)$ with some true underlying parameter value w^*; the value of σ^2 is the same as in our model.

- What is the maximum likelihood estimate of w^* as a function of the training data?

Answer: We know that $y/x \sim \mathcal{N}(w^*, \sigma^2)$. We can therefore estimate w^* by interpreting y_i/x_i as observations. The MLE of w^* is simply the mean

$$\widehat{w^*} = \frac{1}{N}\sum_{i=1}^{N} y_i/x_i.$$

- What is the variance of this estimator due to the noise in the target outputs as a function of N and σ^2 for fixed inputs $\left\{x^i\right\}_{i=1}^N$?

Answer: The variance of $\widehat{w^*}$is simply

$$\begin{aligned} Var\left[\widehat{w^*}\right] &= \frac{1}{N^2}\sum_{i=1}^{N} Var\left[y_i/x_i\right] \ (y_i \text{ are independent}) \\ &= \frac{\sigma^2}{N} \end{aligned}$$