# Statistical Learning Exam

Friday, January 11th 2013, 10:00-13:00

For each question, you can get up to 2.5 points (10 points in total). Do not spend too much time on questions you do not understand! It is wiser to first skip these and start with the questions you find more easy. **Please motivate all your answers!**

1. *Linear Classification* Let $Y_i \in \{-1, 1\}$ and let the $X_i$ be $p$-dimensional vectors of categorical (discrete) or real-valued attributes. Answer the questions below.

   a. Suppose we fit logistic regression to the data set $(X_1, Y_1), \ldots, (X_N, Y_N)$. Can it happen that the resulting coefficient vector $\beta$ will be infinitely large (i.e. $\|\beta\| = \infty$)? If yes, then when does it happen?

   b. Answer the same question as above given that we fit penalized logistic regression (with $L_2$ or $L_1$ regularization).

   c. Suppose we fit logistic regression (not penalized) to the data set $(X_1, Y_1), \ldots, (X_N, Y_N)$. Can it happen that the resulting coefficient vector $\beta$ will be 0? If yes, then when does it happen?

   d. Is it true that fitting logistic regression (not penalized) will result in the decision boundary with the smallest 0/1 training error?

   e. Is it true, that for every training set $(X_1, Y_1), \ldots, (X_N, Y_N)$ that is linearly separable, the decision boundary learned by logistic regression for that training set perfectly separates the data?

   f. Is it true, that for every training set $(X_1, Y_1), \ldots, (X_N, Y_N)$ that is linearly separable, the decision boundary learned by using the Naive Bayes model perfectly separates the data (naive Bayes is used here with a multinomial model for the categorical attributes, and a Gaussian model with fixed variance for the real-valued attributes)?

2. *K-means*. The $K$-means algorithm iteratively minimizes the $W(C)$ criterion, where $C$ is the clustering (function, which assigns to each observation a cluster index), while $W(C)$ is the within-point scatter (within-cluster variance). Please answer the following questions. Motivate your answer in each case.

   a. Is $W(C)$ decreasing after each iteration?

   b. The $K$-means starts by choosing the initial cluster centroids as randomly selected $K$ points. If we start from a different initial centroids, will we end up with a possibly different final solution?

   c. The $K$-means starts by choosing the initial cluster centroids as randomly selected $K$ points. Consider the data set presented on Figure 1. Each dot represents a single point ($p = 2$, so the data can be visualized on a plane). The algorithm is run with $K = 3$, and the initial 3 cluster centroids are marked as white dots in the upper-right cluster. Can you sketch how will the situation look like after the first iteration of $K$-means.

   d. For a given $K$, let $C_K^*$ denote the optimal cluster assignments, minimizing $W(C)$. How would a plot of $W(C_K^*)$ look like as a function of $K = 1, 2, 3, \ldots$? (do not try to calculate particular values, just try to anticipate the shape of the function).

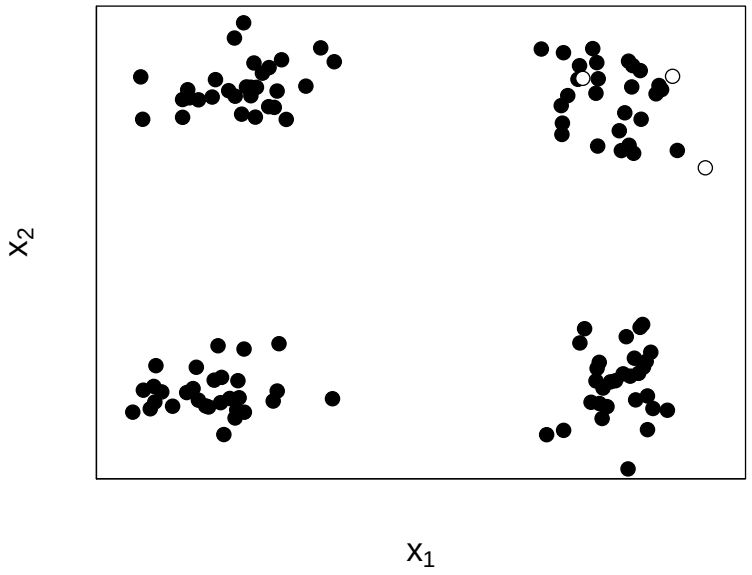


Figure 1: Clustering Problem

3. *Regression*. The least squares regression is the most popular, but not the only way to fit a model to data with continuous output. Another important algorithm is the *least absolute deviation* (LAD) method, which is based on minimizing the sum of absolute errors:

$$SAE(\beta) = \sum_{i=1}^{N} |y_i - x_i^T \beta|.$$

Comparing to least squares, the only difference is that the error is measured as the absolute difference, rather than the squared difference. The main drawback of LAD is that there is no closed-form solution and one needs to numerically solve for $\beta$.

Please answer the questions below. Motivate your answer in each case.

a. The least squares method is known to be equivalent to maximum likelihood estimation, when the noise distribution is Gaussian. Can you propose a distribution, for which $\hat{\beta}$ obtained from LAD would also be a maximum likelihood estimator with respect to this distribution? (you do not need to calculate normalization constant, you can leave the distribution unnormalized, that is you can write "$p(y|x) \propto$ *something*", where $\propto$ means "proportional to").

b. Consider a data set concerning a one-dimensional ($p = 1$) regression problem. Here $x_i = (1, x_{i1})$ and $\beta = (\beta_0, \beta_1)$. The set is presented on Figure 2a (left plot). Input values ($x_{i1}$) are on the horizontal axis, while output values ($y_i$) are on the vertical axis. Can you sketch, what would be the results of fitting:

- least squares,
- least absolute deviation

to the data set? Do not try to calculate $\hat{\beta}$, just describe what you expect to obtain. Figure 2b (right plot) should be helpful for you: look how the two error curves differ for large deviations $y - \hat{y}$.

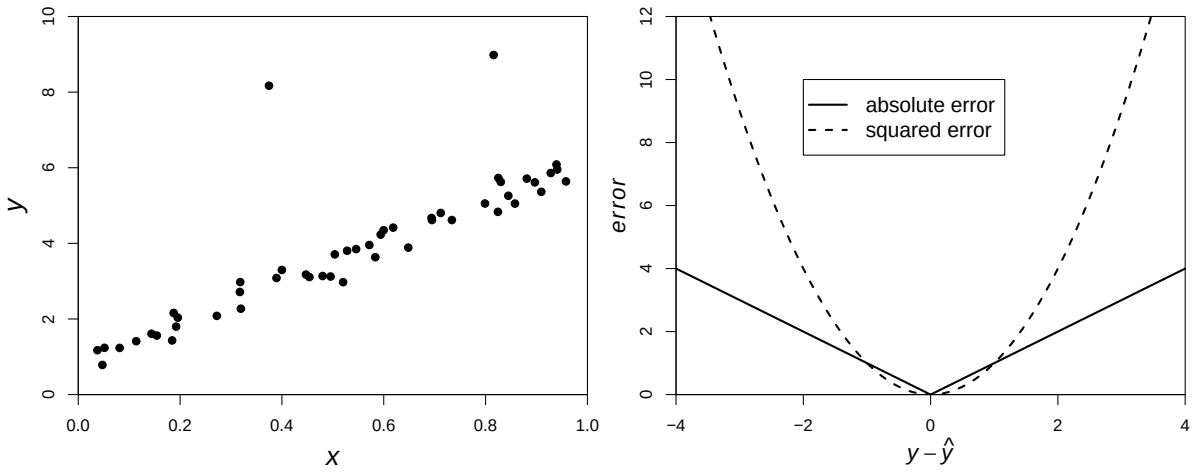


Figure 2: (a) Left plot: regression problem (**pay attention to the two points far above the rest!**). (b) Right plot: the values of squared error $(y-\hat{y})^2$ and absolute error $|y-\hat{y}|$ as a function of $y-\hat{y}$, where $y$ is the observed output and $\hat{y}$ is the output from the model (model prediction).

c. The least-squares estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ has the property that, if all $x_i$ are 0, then $\hat{\beta}_0$ is always well-defined, and equal to the average of the $y_i$'s. Give an example of a nonempty data set with all $x_i$ equal to 0 for which the LAD estimator is not defined (i.e. there are various values of $\beta$ minimizing $SAE(\beta)$). Also give an example of such a data set for which the LAD estimator is uniquely defined, but *not* equal to the average of the $y_i$ (HINT: for the first case there exists an example with only two data points, for the second case, there exists an example with only three data points. Just try out a few values for the $y_i$ points and a few values of $\beta$ and calculate the corresponding $SAE(\beta)$ to get an idea of how $SAE(\beta)$ varies with $\beta$).

d. Suppose we have a data set (now the $x_i$ do not have to be 0 any more) with least-squares estimate $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ and LAD estimate $\beta^\circ = (\beta_0^\circ, \beta_1^\circ)$. We translate all the $y$-values in our data set by adding, say, 2 to them and we recalculate the least squares and LAD estimator. What will happen to $\hat{\beta}_0, \hat{\beta}_1, \beta_0^\circ, \beta_1^\circ$? (You don't have to formally prove that your answer is correct, just give some intuition on why you think your answer is correct).

4. *Classification with a Strange Model.* Suppose we observe a sample $(x_1, y_1), \ldots, (x_N, y_N)$ with each $y_i \in \mathcal{Y} = \{0, 1, \ldots, 9\}$ (there are 10 classes) and each $x_i$ a positive integer, i.e. $x_i \in \mathcal{X} = \{1, 2, \ldots\}$. The goal is to make good predictions of a new $Y \in \mathcal{Y}$ given a new $X \in \mathcal{X}$. In the exercise below you may assume that $N$ is reasonably large, at least, say 1000.

   The company `wefititall.com` proposes to fit the data using the following unusual single-parameter model $\mathcal{F} = \{f_\theta : 0 \leq \theta < 1\}$ consisting of functions $f_\theta : \mathcal{X} \to \mathcal{Y}$ defined as follows:

   $$f_\theta(x) = \text{the } x\text{-th digit of } \theta.$$

   For example, if $f_\theta(1) = 3, f_\theta(2) = 1, f_\theta(3) = 4$, $f_\theta(4) = 1$, then $\theta = .3141...$, i.e. its first four digits must be $3, 1, 4, 1$. As another example, for $\theta = 1/3 = 0.3333...$ we have $f_\theta(x) = 3$ for all $x$.

   a. Suppose we fit this model using empirical error minimization with the 0/1-loss, i.e., we

pick the $\hat{\theta}$ minimizing $\hat{L} = \sum_{i=1}^{N} L(y_i, f_\theta(x_i))$ where $L(y_i, \hat{y}_i) = 0$ if $y_i = \hat{y}_i$ and 1 otherwise. If the minimum $\hat{K}$ is achieved for more than one $\theta$ with $0 \leq \theta < 1$, we pick the smallest one. (1) What $\hat{\theta}$ will be chosen? (2) Assume there are no repetitions in the $x_i$. How large will its error $\hat{L}$ on the training set be?

b. Suppose the data $(X_i, Y_i)$ are i.i.d. according to some unknown distribution $P^*$, and we test $\hat{\theta}$ on a new example $(X, Y)$ drawn from the same distribution $P^*$. What can you say about the expected prediction error (EPE) of $\hat{\theta}$ as compared to its empirical error $\hat{L}$?

c. It is sometimes claimed that empirical error minimization is a good method, that will not overfit, as long as the number of parameters is small relative to the sample size $N$. Does this claim hold in general? Why (not)?

d. Let $\Theta_k$ be the subset of $[0, 1]$ consisting of all numbers with the $k+1$st, $k+2$nd etc. digit after the decimal point being 0. For example, $\Theta_2 = \{0, 1/100, 2/100, \ldots, 99/100\} = \{0, 0.01, 0.02, \ldots, 0.99\}$. Suppose that, based on data as above, we select a submodel $\Theta_{\hat{k}}$ where $\hat{k}$ is chosen using 10-fold cross-validation. Let $\hat{\theta}_{\hat{k}}$ be the $\theta$ that minimizes the training error $\hat{L}_{\hat{k}}$ within $\Theta_{\hat{k}}$. What can you say about the EPE of $\hat{\theta}_{\hat{k}}$ as compared to its empirical error $\hat{L}_{\hat{k}}$?