RPZ **January 9, 2015**

Name:

1. **(7 points)** A die is cast independently a number of times until a six appears on the up side. The probability such event happens after n casts is:

$$p(n) = (1-q)^{n-1}q,$$

where q is un unknown probability of six showing on the up side.

   (a) Explain the form (geometric distribution) of $p(n)$. **(1 point)**

   (b) Derive the maximum likelihood estimate for q. **(3 points)**

   (c) Let's assume that an upper limit on the number of casts n_{max} has been set. What conclusion can be made about q if n_{max} is reached, i.e. if $n > n_{max}$? **(1 point)**.

   (d) What is the maximum likelihood estimate of q if you were just told that in n_{max} casts six never showed on the up side? **(1 point)**

   (e) Let's assume that it is know a priori that $q \in [0.1, 0.2]$. What would be the maximum likelihood estimate in this case? Consider e.g. the case when six shows up in the first cast. **(1 point)**

2. **(5 points)** Define the minimax decision problem **(2 points)**. Find the optimal minimax strategy in the following two-class problem: the observations are real numbers $x \in [0, 1]$. Conditional probability densities $p(x|k)$ are $p(x|1) = x/2$ and $p(x|2) = 6x(1-x)$. **(3 points)**

3. **(7 points)**. K-means. Describe the basic algorithm **(2 points)**.

   Consider the following problem. K wells can be drilled at any location (X_k, Y_k) at fixed cost C_w per well to provide water to n huts in positions (x_i, y_i). A pipe can be laid in a straight line between a well and a hut at a cost C_p per meter.

   (a) Define the cost of the installation in terms of the number of wells, K, well locations, $\{(X_k, Y_k)\}_{k=1}^{K}$, and hut positions, $\{(x_i, y_i)\}_{i=1}^{n}$. **(1 point)**

   (b) Modify the K-means algorithm so that it minimizes the cost of drilling wells and laying the pipes, i.e. so that it finds a local minimum of the total cost of installation. **(2 points)**

   (c) Explain how to set K. **(1 point)**

   (d) Consider the case when pipes can be laid only in north-south and east-west directions along existing roads. Modify the K-means accordingly. **(1 point)**

4. **(6 points)**. Describe the Support Vector Machine, its learning method and its properties **(3 points)**, including the way it handles non-separable training data **(1 point)**. Discuss the use of kernels and the differences between linear and kernel SVM in learning and in classification **(2 points)**.