Solution Sketches
Midterm Exam
COSC 6342 *Machine Learning*
March 20, 2013

Your Name:
Your student id:

Problem 1 [5+?]: Hypothesis Classes
Problem 2 [8]: Losses and Risks
Problem 3 [11]: Model Generation
Problem 4 [8]: Parametric Model Generation/Mahalanobis
Problem 5 [8]: EM (and K-means)
Problem 6 [8]: DBSCAN (and K-means)
Problem 7 [4]: Non-Parametric Prediction Approaches
Problem 8 [8]: General Questions

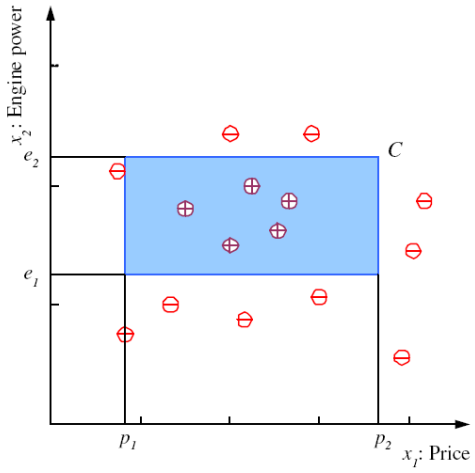
Σ [60]:

Grade:

The exam is "open books and notes" and you have 75 minutes to complete the exam. The exam will count approx. 28% towards the course grade.

1) Hypothesis Classes [5]

Assume the hypothesis class for family car problem, discussed in Chapter 2 of the textbook, is a triangle. What are the parameters of the hypothesis class? Describe an approach that calculates the parameters of the triangle from a set of examples—more sophisticated approaches will receive extra credit up to 3 points [5+[3]]



Parameters: 3 points (6 numbers in 2D) as a triangle is defined by 3 points.
No algorithm given!

2) Losses and Risks [8]

Determine the optimal decision making strategy with reject option for the following cancer diagnosis problem [6]!

C1=has cancer

C2=has not cancer

$\lambda_{12}=0.2$, $\lambda_{21}=1$ (as "always" we assume: $\lambda_{11}=0$, $\lambda_{22}=0$)

$\lambda_{reject,2}=0.1$

$\lambda_{reject,1}=0.5$

Inputs: P(C1|x), P(C2|x)

Decision Making Strategy: ...

The risks associated with the 3 options are as follows:

R(a1|x) = 0.2P(C2|x)

R(a2|x) = P(C1|x)

R(reject|x) = 0.5P(C1|x)+0.1P(C2|x)

Equating all 3 pairs of risks[1], we find out that all three risks are equal if P(C1|x)=1/6. After analyzing which risk it higher if P(C1|x) is greater/smaller 1/6 we obtain the following decision rule:

[1] For example, setting P(C1|x)=0.5P(C1|x)+0.1(1-P(C1|x)) we obtain 0.6P(C1|x)=0.1 therefore P(C1|x)=1/6; equating the other two pairs of risk functions leads to the same result!

If P(C1|x) > 1/6, choose class 1
If P(C1|x) < 1/6, choose class 2
If P(C1|x) = 1/6, choose class1 or class2 or reject[2].

Summarize when the decision C1="Has Cancer" will be taken in your strategy. [2]
Hence tell the patient she/he has cancer if P(C1|x) > 1/6

3) Model Generation [11]

a) Assume you have a model with a high bias and a low variance. What are the characteristics of such a model? [2]
One answer is suggesting underfitting!
Another answer: the model is simple (therefore high bias, as the model is too simple to obtain a good match with the distribution in the dataset) and can be learnt (due its simplicity easily) just using a few training example and is not very sensitive to noise (therefore low variance)

b) Assume you have a small dataset from which a model has to be generated. Would you prefer to learn a complex model or a simple model in this case? Give reasons for your answer! [3]
The simple model should be preferred. Reason: it will not be possible to learn the good model due to the small number of examples—particularly the model's variance will be high leading to a high generalization error.

c) What is overfitting? Limit your answer to 2-3 sentences! [3]
Definition:
+ Let D is the training set, D' is the testing set. We say a hypothesis h is overfitting in a dataset D if there is another hypothesis h' in the hypothesis space where h has better classification accuracy than h' on D but worse classification accuracy than h' on D'.
Or:
+ Overfitting is the phenomenon in which: when the number of parameters increases (the model gets more complex), the training error decreases but the testing error increases.
- Low bias and high variance ⇨ overfitting.

d) What objective function does maximum likelihood estimation minimize in parametric density estimation? How does the maximum likelihood approach differ from the maximum a posteriori (MAP) approach? [3]
- MLE maximizes the $\ln P(D|\theta)$ likelihood of model (θ) with respect to data D.
- MAP maximizes the $\ln P(\theta|D)$, likelihood of data D with respect to model θ.

Moreover, MAP additionally uses priors.

[2] All three risks are the same in this case!

4) Parametric Model Generation / Mahalanobis Distance [8]

a) Assume we have a dataset with 3 attributes with the following covariance matrix:

1	0	-1
0	1	0
-1	0	1

What does this co-variance matrix tell you about the relationship of the three attributes? [2]

- Three attributes have the same variance.
- Attributes 1 and 2 are independent/there is no linear relationship between the attributes.
- Attributes 3 and 2 are independent/there is no linear relationship between the attributes.
- Attributes 1 and 3 have negative correlation of -1 (attribute1=−λ *attribute2)

b) What are the advantages of using Mahalanobis distance over Euclidean distance? [2]

- Malanobis distance normalizes attributes based on variance; this makes all independent attributes equally important, and
- it alleviates problem caused by using different scales, and
- downplays the contribution of correlated attributes in distance computations

c) For what problems does parametric model estimation **not** work well? [4]

Parametric model estimation does not work well in problems which have:

- The dataset does not match well the assumption of the employed parametric model estimation technique.
- Feature space has high number of dimensions.
- If there is no global model and only regional/local models exist.
- Parametric models are kind of simplistic and therefore might not capture the characteristic of the data.

5) EM and K-means [8]

a) EM uses a mixture of *k* Gaussian for clustering; what purpose do the k Gaussian serve? [2]

k Gaussians serve as a model for the k clusters—one Gaussian per cluster---each Gaussian is used to compute the probability that an object belongs to a particular cluster.

b) What is the task of the E-step of the EM-algorithm? Give a verbal description (and not (just) formulas) how EM accomplishes the task of the E-step! [4]

E-step is Expectation step. Compute the posterior distribution

$$Q(\theta', \theta^{(t-1)}) = E_{T|Z,\theta^{(t-1)}}[l_0(\theta'; T)]$$

The E-step computes for each object o the probability that it belongs to each of the *k* clusters$_1$,…,clusters$_k$. [2.5] The step is done by dividing the density of P(C_i|o) for i=1,..,k by the sum of the densities of o with respect to the k Gaussians.

c) Characterize what the E-step of K-means does! [2]
It forms clusters by assigning each object o in the dataset to the cluster with the closest centroid to o.

6) DBSCAN and K-means [8]

a) What is a core point in DBSCAN? What role do core points play in forming clusters? [3]
A point is a core point if it has more than a specified number of points (MinPts) within Eps
DBSCAN takes an unprocessed core point p and using this core-point it forms a cluster that contains all core- and border points that are density-reachable from p; this process continues until all core-points have been assigned to clusters. In summary, forms clusters by recursively computing points in the radius of core points.

b) Compare DBSCAN and K-means; what are the main differences between the 2 clustering approaches? [5]
- DBSCAN has the potential to find arbitrary shape clusters, whereas kmeans is limited to clusters that take the shape of convex polygons [2]
- DBSCAN has outlier detection, but not kmeans [1]
- K-Means performs an iterative maximization procedure, whereas DBSCAN forms clusters in a single iteration [1]
- K-means results depend on initialization and different clusters are usually obtained for different intializations; DBSCAN is more or less deterministic (the only exception is the assignment of borderpoints that lie in the radius of multiple core points)[1]
- K-means is basically O(n), DBSCAN is O(n*log(n))/O(n**2) [1]
At most 5 points, even if all 5 answers are given!

7) Non-parametric Prediction and Traditional Regression [4]

Compare traditional regression with non-parametric prediction approaches, such as regressograms/kernel smoothers? What are the main differences between the two approaches?

1. Traditional regression approaches generate a **single** model[1] that is computed **using all examples** of the training set[0.5].
2. Non-parametric approaches employ **multiple (local) models** [0.5] that are derived from a **subset of the training examples in the neighborhood of the query point**[3] . [1]
3. Non-parametric approaches are lazy, as they create the model on the fly, and not beforehand as the traditional regression approach does. [1]

[3] **or by using a weighted sampling approach which assigns higher weights to points that are closer to the query point.**

8) General Questions [8]

a) What is the goal of PCA? Limit your answer to 3-4 sentences! [3]

Your answer should mention:

"PCA seeks for linear transformations to a lower dimensional space, and it tries to capture most of the variance of data…"

b) Most decision tree learning algorithms, such as C4.5, construct decision trees top-down using a greedy algorithm—why is this approach so popular? [3]

As learning the "optimal" decision trees is NP-hard (very time consuming), it is not realistic to find the optimal decision tree.[1.5] Greedy algorithms, as they reach solutions quickly without any backtracking, make it feasible to create a "decent", but not optimal decision tree relatively quickly.[1.5]

c) What objective function do regression trees minimize? [2]

It computes the variance of the values of the dependent variable (output variable) of the objects that are associated with the node of a tree; tests that lead to a lower overall variance are preferred by the regression tree induction algorithm.