UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK4030 — Statistical Learning: Advanced Regression and Classification
Day of examination:	Monday, December 11th, 2017
Examination hours:	9.00 – 13.00

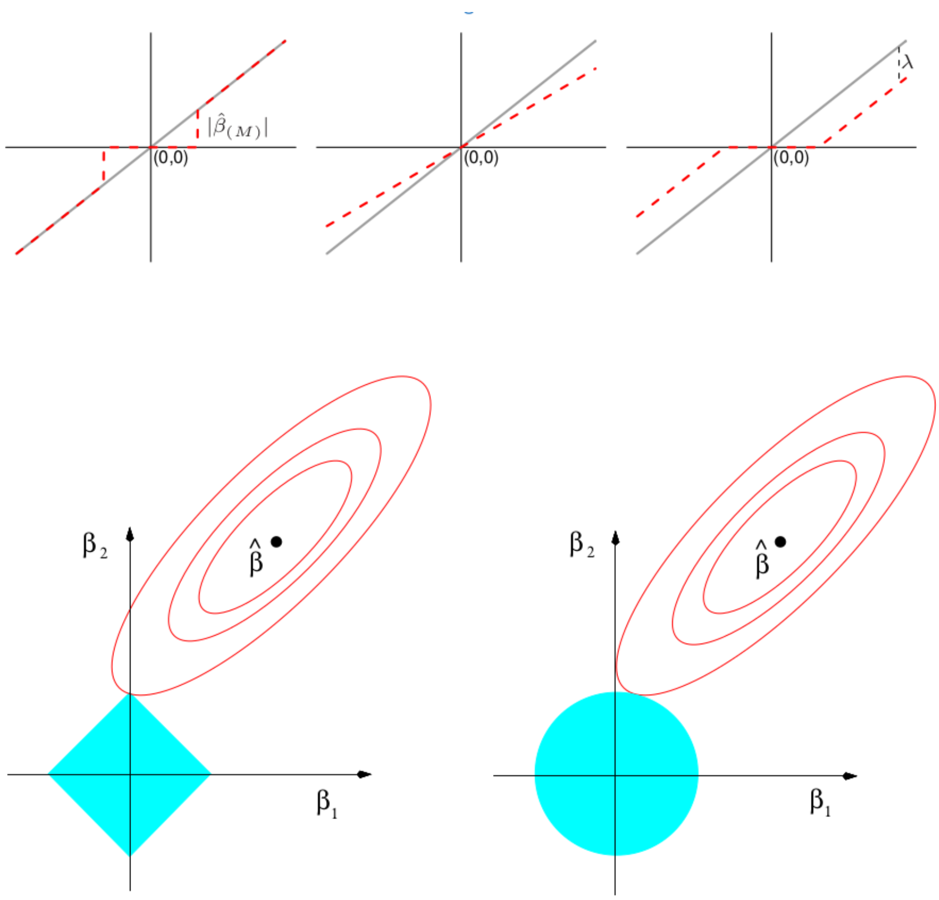
This problem set consists of 3 pages.

Appendices:	None.
Permitted aids:	None.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1 Penalized regression

Consider the following figure from the text book (Hastie, Tibshirani & Friedman, 2009, The Elements of Statistical Learning):



(Continued on page 2.)

a

Identify the techniques that the three top plots refer to and explain what these plots show.

b

Explain what it is shown in the bottom two plots.

c

In the context of linear regression, show analytically that the ridge estimator has larger bias than the ordinary least square estimator.

d

In the context of linear regression, show analytically that the ridge estimator has smaller variance than the ordinary least square estimator.

Problem 2 Cross-validation

Consider a classification problem in which a large number of continuous predictors is available. The following procedure is applied:

1. reduce the number of predictors by selecting only those that are most correlated with the outcome;
2. build a multivariate classifier using the predictors selected in the previous step;
3. use cross-validation to estimate the unknown tuning parameters and to estimate the prediction error of the final model.

a

Explain why this procedure is incorrect.

b

Suggest an alternative procedure to derive a correct estimate of the prediction error of the final model.

Problem 3 Boosting

Consider the AdaBoost algorithm for classification, where the outcome is $y \in \{-1; 1\}$:

(Continued on page 3.)

Algorithm 10.1 *AdaBoost.M1.*

1. Initialize the observation weights $w_i = 1/N,\ i = 1, 2, \ldots, N$.
2. For $m = 1$ to M:
   (a) Fit a classifier $G_m(x)$ to the training data using weights w_i.
   (b) Compute
   $$\text{err}_m = \frac{\sum_{i=1}^N w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i}.$$
   (c) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$.
   (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))],\ i = 1, 2, \ldots, N$.
3. Output $G(x) = \text{sign}\left[\sum_{m=1}^M \alpha_m G_m(x)\right]$.

a

Describe the original idea behind the AdaBoost algorithm.

b-c-d

Show that the AdaBoost algorithm reported above can be interpreted as a forward stagewise modelling procedure which minimizes the loss function $L(y, f(x)) = \exp\{-yf(x)\}$. Following this interpretation, the current estimate $f_{m-1}(x)$ is updated by adding the step-specific result of the classifier $G_m(x_i)$ to produce a new estimate $f_m(x)$. In particular, at each step m one must find G_m and α_m such that

$$(\alpha_m, G_m) = \text{argmin}_{\alpha,G} \sum_{i=1}^N \exp\{-y_i[f_{m-1}(x_i) + \frac{\alpha}{2}G(x_i)]\}.$$

Show that:

b

$$G_m(x) = \text{argmin}_G \sum_{i=1}^N w_i^{(m)} I(y_i \neq G(x_i)),$$

where $w_i^{(m)} = \exp\{-y_i f_{m-1}(x_i)\}$;

c

$$\alpha_m = \log \frac{1 - \text{err}_m}{\text{err}_m},$$

where $\text{err}_m = \frac{\sum_{i=1}^N w_i^{(m)} I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i^{(m)}}$;

d

$$w_i^{(m+1)} \propto w_i^{(m)} \exp\{\alpha_m I(y_i \neq G_m(x_i))\}.$$

THE END