UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK-IN4300/STK-IN9300 — Statistical learning methods in Data Science
Day of examination:	Thursday, December 5th, 2019
Examination hours:	14.30 – 18.30

This problem set consists of 4 pages.

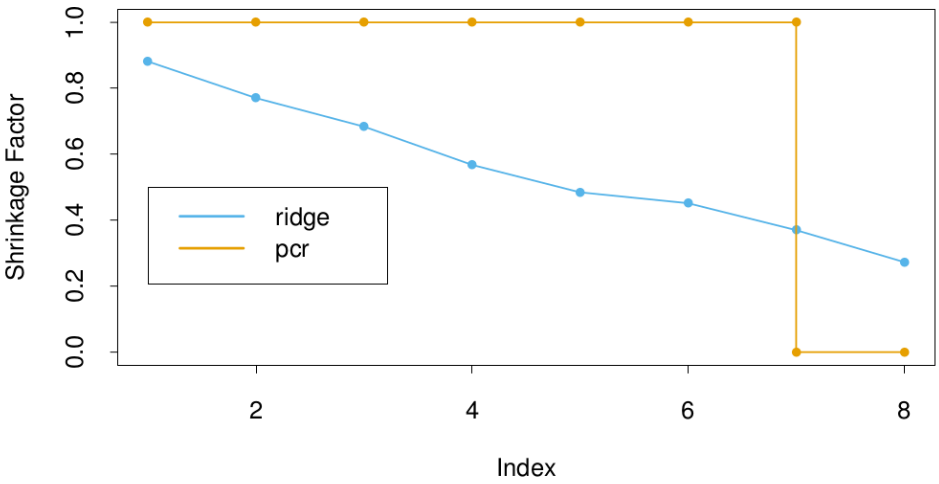
Appendices:	None.
Permitted aids:	None.

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

Problem 1 Penalized regression

a Ridge versus principal component regression (10 pt.)

Consider the following figure from the textbook (Hastie, Tibshirani & Friedman, 2009, The Elements of Statistical Learning, Figure 3.17), where the x-label "Index" denotes the index of the principal components:



Explain the figure above, highlighting the differences between *ridge regression* and *principal component regression* when it concerns their shrinkage effect.

(Continued on page 2.)

b Sparse group lasso

Consider the following version of lasso, called *sparse group lasso*,

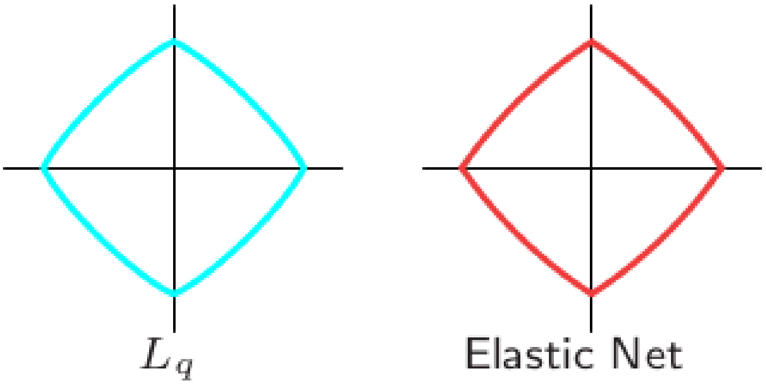
$$\min_\beta \left\{ \left|\left|(y - \beta_0 \vec{1} - \sum_{\ell=1}^{L} X_\ell \beta_\ell \right|\right|_2^2 + (1-\alpha)\lambda \sum_{\ell=1}^{L} \sqrt{p_\ell}\, ||\beta_j||_2 + \alpha\lambda\, ||\beta||_1 \right\},$$

where $\vec{1}$ denotes an N-dimensional vector of 1s, $\lambda \geq 0$ and $0 \leq \alpha \leq 1$. Answer to the following questions:

- Why does β_0 appears only in the first term? **(3 pt.)**
- What does it happen when $\alpha = 0$ and $\alpha = 1$, respectively? **(2 pt.)**
- Briefly describe the concept of "bet on sparsity". **(5 pt.)**

c Elastic net versus bridge regression (10 pt.)

Briefly describe *elastic net* and *bridge regression*, and explain why, despite the corresponding constraints are almost indistinguishable in the figure here below (Hastie, Tibshirani & Friedman, 2009, The Elements of Statistical Learning, Figure 3.13), they provide, in general, quite different models.



Problem 2 Ensemble Methods

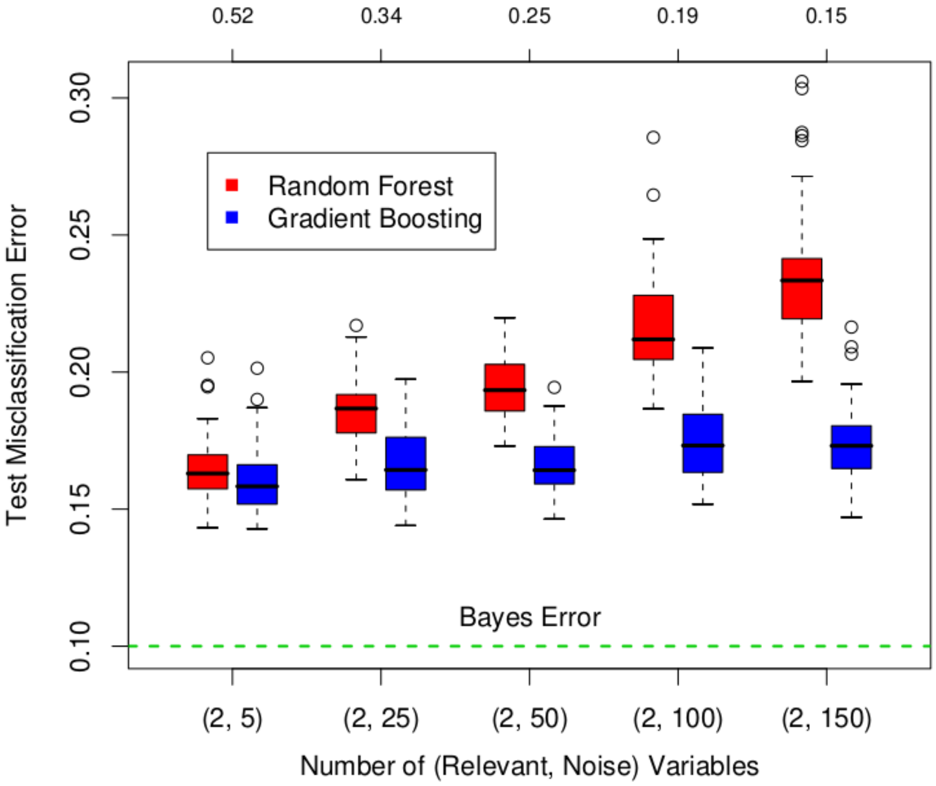
a Bagging (10 pt.)

Consider a classification problem and how to aggregate the results of the single trees in a bagging classifier. The aggregation can be done by looking at the estimated classes or at the class-probability estimates. Show with a simple example that the two procedures can produce different results in terms of classification of an observation.

b Random Forests (10 pt.)

Consider the figure below (Hastie, Tibshirani & Friedman, 2009, The Elements of Statistical Learning, Figure 15.7),

(Continued on page 3.)



in which the results over 50 simulations (each time a training set of 300 observations has been generated, together with a test set of 500 observations) for *random forests* (red box-plot) and *gradient boosting* (blue box-plots, not relevant for the exercise) have been reported. Here the true boundary (it is a binary classification problem) depends on two variables, and an increasing number of noise variables are added (see x-axis). The default $m = \sqrt{p}$, where m is the number of candidate variables randomly selected as input before each split in the trees and p is the total number of variables, has been used. Explain why the performance of *random forests* worsen with p increasing.

c Boosting 1 (10 pt.)

Consider the regression model $y_i = f(x_i) + \epsilon_i, i = 1, \dots, N$, where ϵ_i are i.i.d. random variables with $E[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$. Bühlmann & Yu (2003, Journal of the American Statistical Society) showed that, for L$_2$Boost, if

$$\frac{\mu_k^2}{\sigma^2} > \frac{1}{(1-\lambda_k)^2} - 1 \qquad (1)$$

for all k with $\lambda_k < 1$, then $\text{MSE}_{\mathcal{B}_m} < \text{MSE}_{\mathcal{S}}$, where $\text{MSE}_{\mathcal{B}_m}$ and $\text{MSE}_{\mathcal{S}}$ denote the mean square errors obtained using the boosting operator and the corresponding linear operator $\mathcal{S}$ used as base learner, respectively. Here, λ_k is the k-th eigenvalue of $\mathcal{S}$ and μ_k represents the true regression function corresponding to the k-th eigenvector of $\mathcal{S}$.

Interpret Equation (1), focusing on the importance of "shrinkage" in boosting.

(Continued on page 4.)

d Boosting 2

Consider the following component-wise gradient boosting algorithm,

```
1. initialize the estimate, e.g., $\hat{f}_j^{[0]}(x) \equiv 0, j = 1, \dots, p$;
2. for $m = 1, \dots, m_{\text{stop}}$,
    • compute the negative gradient vector, $u = -\left.\frac{\partial L(y,f(x))}{\partial f(x)}\right|_{f(x)=\hat{f}^{[m-1]}(x)}$;
    • $\forall j$, fit the base learner to the negative gradient vector, $\hat{h}(u, x_j)$;
    • select the best update $j^*$;
    • update the estimate, $\hat{f}_{j^*}^{[m]}(x) = \hat{f}_{j^*}^{[m-1]} + \nu \hat{h}(u, x_{j^*})$;
3. final estimate, $\hat{f}_{m_{\text{stop}}}(x) = \sum_{j=1}^{p} \hat{f}_j^{[m_{\text{stop}}]}(x)$.
```

where $L(y, f(x))$ is a generic loss function and $h(\cdot)$ a base-learner:

- Identify the tuning parameters of the algorithm; **(2 pt.)**
- Describe how they are usually computed in practice; **(2 pt.)**
- Relate them to the prediction performance of the final model; **(3 pt.)**
- Explain why their optimal values are related to each other. **(3 pt.)**

Problem 3 Bias-variance trade-off

a Expected prediction error (10 pt.)

Consider $y = f(x) + \epsilon$, with $\mathrm{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \sigma_\epsilon$. Show mathematically that, in the case of squared-error loss, the expected prediction error of a regression fit $\hat{f}(x)$ at an input point $x = x_0$ can be decomposed into: irreducible error, squared bias, variance. Moreover, briefly explain what these three terms are.

b Boosting (10 pt.)

Consider the model $y_i = f(x_i) + \epsilon_i$, $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{Var}[\epsilon_i] = \sigma$, $i = 1 \dots, N$. Derive the formula of the squared bias for L_2Boost,

$$\mathrm{bias}(m, \mathcal{S}; f)^2 = N^{-1} f^T U \mathrm{diag}((1 - \lambda_k)^{2m+2}) U^T f,$$

when a symmetric learner $\mathcal{S}$, with eigenvalues λ_k and eigenvectors building the columns of the orthonormal matrix U, is used. Here f denotes the vector of the true regression function and m the number of boosting steps.

Hint: remember that the L_2Boost operator $\mathcal{B}_m$ can be rewritten as $\mathcal{B}_m = U D_m U^T$, with $D_m = \mathrm{diag}(1 - (1 - \lambda_k)^{m+1})$ and $UU^T = U^T U = I$.

c Model complexity (10 pt.)

Relate the concept of model complexity to the concept of bias-variance trade-off, and show how this works for the k-nearest neighbours algorithm.

THE END