UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK-2100 — Machine Learning and Statistical Methods for Prediction and Classification
Day of examination:	Friday 31st of May
Examination hours:	15.00 – 19.00
This problem set consists of 6 pages.	
Appendices:	None.
Permitted aids:	Approved calculator and List of formulas for STK1100/STK1110

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

Problem 1 kNN

a

Consider the k nearest neighbours method with $k = 1$, let us call it 1-NN. Explain why its training error is always 0, and use this example to illustrate the concept of overfitting.

b

We want to use our 1-NN method to estimate the value of the origin. When the points are uniformly distributed in the p-dimensional space, we now that the median distance between the origin and the closest point is

$$d(p, n) = \left(1 - \frac{1}{2}^{\frac{1}{n}}\right)^{\frac{1}{p}},$$

with n being the sample size. Use this formula to illustrate the concept of "curse of dimensionality".

c

Let us tune the parameter k. After having briefly described the 0.632 bootstrap approach, explain why it may not be a good idea to use it to tune k, especially if the optimal k is small. Suggest an alternative.

Problem 2 Regression

It is known that a drawback of regression trees is their large variability. To mitigate this issue, bagging and random forests have been developed.

(Continued on page 2.)

a

Assume the training observations (x_i, y_i), $i = 1, \ldots, n$, independently drawn from a distribution $\mathcal{P}$. Define the "ideal" bagging estimator $f_{\text{ag}}(x) = E_{\mathcal{P}}[f^*(x)]$, where $f^*(x)$ is an estimator, let us say a tree, based on the "bootstrap sample" (x_i^*, y_i^*), $i = 1, \ldots, n$ sampled from $\mathcal{P}$.

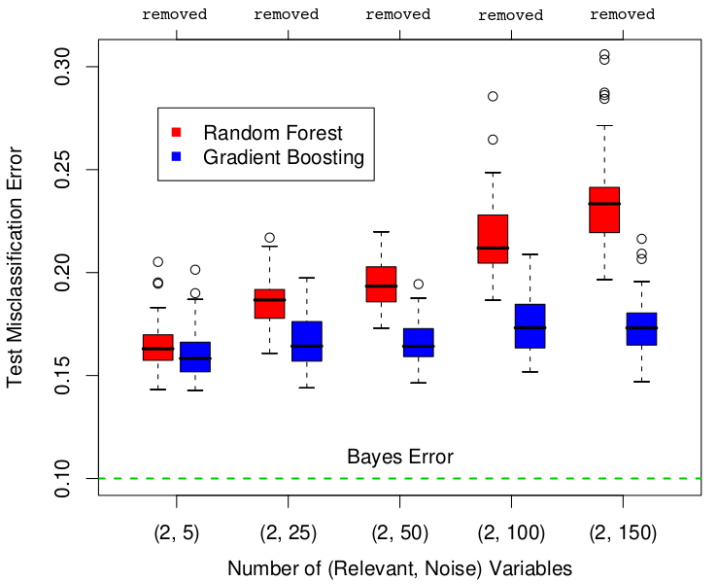
Show mathematically that the mean square error loss of an "ideal" bagging estimator (i.e., the "population aggregator" defined above) is always smaller than that of $f^*(x)$.

b

Describe the Random Forests method in the form of an algorithm, highlighting the step that makes the Random Forests "better" (in terms of reduced variance) than bagging.

c

Consider (a modified version of) Figure 15.7 from the textbook (Elements of Statistical Learning by Hastie, Tibshirani, and Friedman),



The figure's caption is *A comparison of random forests and gradient boosting on problems with increasing numbers of noise variables. In each case the true decision boundary depends on two variables, and an increasing number of noise variables are included. Random forests uses its default value* $m = \sqrt{p}$. *At the top of each pair is the probability that one of the relevant variables is chosen at any split* [here removed] . *The results are based on 50 simulations for each pair, with a training sample of 300, and a test sample of 500.*

Based on this figure, explain why the performance of Random Forests decreases when the number of noisy variables increases. Moreover, compute the removed values on the top of the plot (hint: the formula for the right distribution is in the List of formulas for STK1100/STK1110).

(Continued on page 3.)

Problem 3 Classification

The Pima Dataset contains information about 768 women of a population, Pima, particularly susceptible to diabetes. The response `diabetes` identifies which of the persons involved in the study developed the disease (`neg` = no, `pos` = yes). Eight continuous independent variables contain information on:

- `pregnant`: number of pregnancies;
- `glucose`: plasma glucose concentration at 2 h in an oral glucose tolerance test;
- `pressure`: diastolic blod pressure (mm Hg);
- `triceps`: triceps skin fold thickness (mm);
- `insulin`: 2-h serum insulin (μU/mL);
- `mass`: body mass index (kg/m^2);
- `pedigree`: diabetes pedigree function;
- `age`: age (years);

a

Consider the following R output of a logistic regression model, let us call it Model A.

```
Call:
glm(formula = diabetes ~ pregnant + glucose, family = binomial,
    data = PimaIndiansDiabetes)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.751540   0.440659 -13.052  < 2e-16 ***
pregnant     0.123287   0.025590   4.818 1.45e-06 ***
glucose      0.037080   0.003275  11.322  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 784.95  on 765  degrees of freedom
AIC: 790.95
```

Based on this model, compute the estimated probability of developing diabetes for a woman who went through 2 pregnancies and has a plasma glucose concentration of 160. How low should the level of plasma glucose concentration be to have a probability of developing the disease less than 0.5?

(Continued on page 4.)

b

When we add other covariates to the model, we obtain the following models:

- Model B

```
Call:
glm(formula = diabetes ~ pregnant + glucose + triceps + age +
    pressure, family = binomial, data = PimaIndiansDiabetes)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.797930   0.520568 -11.138  < 2e-16 ***
pregnant     0.110836   0.030391   3.647 0.000265 ***
glucose      0.036471   0.003371  10.818  < 2e-16 ***
triceps      0.012646   0.005739   2.204 0.027545 *
age          0.013166   0.009039   1.457 0.145224
pressure    -0.007709   0.004729  -1.630 0.103064
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 777.95  on 762  degrees of freedom
AIC: 789.95
```

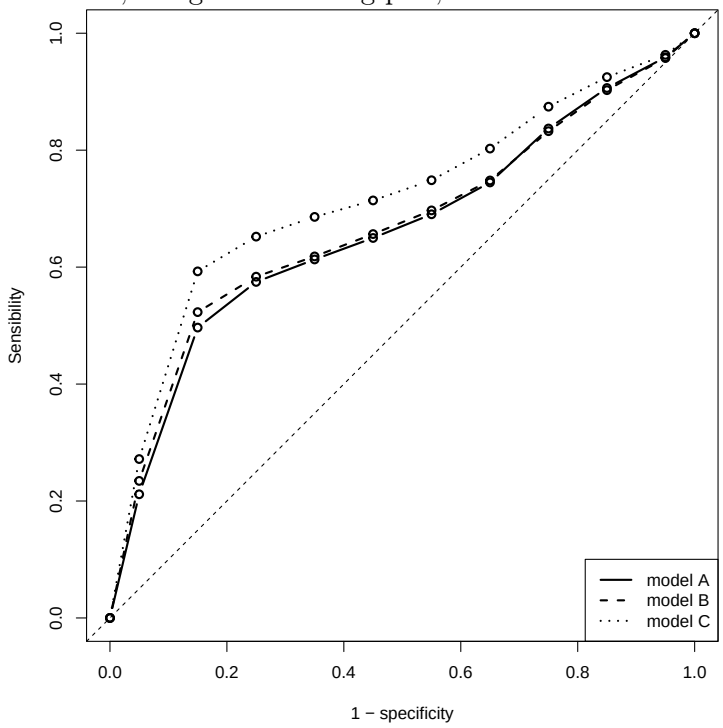
- Model C

```
Call:
glm(formula = diabetes ~ ., family = binomial, data = PimaIndiansDiabetes)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.4046964  0.7166359 -11.728  < 2e-16 ***
pregnant     0.1231823  0.0320776   3.840 0.000123 ***
glucose      0.0351637  0.0037087   9.481  < 2e-16 ***
pressure    -0.0132955  0.0052336  -2.540 0.011072 *
triceps      0.0006190  0.0068994   0.090 0.928515
insulin     -0.0011917  0.0009012  -1.322 0.186065
mass         0.0897010  0.0150876   5.945 2.76e-09 ***
pedigree     0.9451797  0.2991475   3.160 0.001580 **
age          0.0148690  0.0093348   1.593 0.111192
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 759  degrees of freedom
AIC: 741.45
```

(Continued on page 5.)

Explain why the covariate `pressure` does not seem to have an effect on the response in model B, but it seems to have an effect in model C.

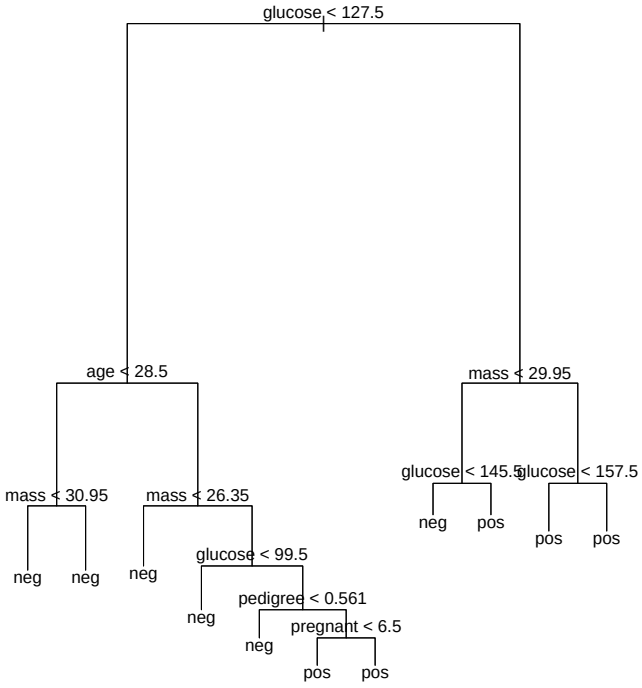
Moreover, using the following plot,



explain which of these three models you would considered the "best" and why. Is any of them really satisfactory? Why?

c

Consider the following classification tree, estimated on the same data,



(Continued on page 6.)

Provide the estimate for the same individual of point **a** (2 pregnancies, `glucose` = 160), supposing that her body mass index is 28 kg/m^2. Indicate how you have found that value.

Moreover, explain why the most left and the most right splits in the tree plot are there, despite both leaves give the same result (both `neg` and both `pos`, respectively).

Problem 4 Clustering

a

Describe the K-means method, including a possible way to find the best K. Why cannot we use cross-validation?

THE END