UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK-IN4300/9300 — Statistical Learning Methods in Data Science
Day of examination:	Tuesday, 28 November 2023
Examination hours:	15.00 – 19.00
This problem set consists of 5 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 Fitting a Regression Function

Consider a regression setting and assume an additive error model:

$$Y = f_\theta(X) + \epsilon,$$

where θ denotes the model parameters defining the regression function f_θ and $\epsilon \sim N(0, \sigma^2)$ is the error term.

(a) For the purpose of estimating θ from data, describe (i) the least squares method and (ii) the maximum likelihood method.

(b) Show that the least squares method and the maximum likelihood method are equivalent under the considered assumptions described above. Hint: recall that the probability density function of a random variable $Z \sim N(\mu, \sigma^2)$ is given by:

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}, \; z \in \mathbb{R} \text{ and } \mu \in \mathbb{R}, \sigma > 0.$$

Problem 2 Bias-Variance Tradeoff

Consider a regression setting with an additive error model $Y = f(X) + \epsilon$, where $\mathbb{E}[\epsilon] = 0$ and $\text{Var}[\epsilon] = \sigma_\epsilon^2$.

(a) Using the squared-error loss, show that the expected prediction error of a regression fit $\hat{f}(X)$ at input point $X = x_0$ can be decomposed as:

$$\mathbb{E}[(Y - \hat{f}(x_0))^2 \,|\, x_0] = \sigma_\epsilon^2 + \big(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\big)^2 + \mathbb{E}\big[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2\big],$$

and describe what the three terms on the right-hand side of the above equation represent. Hint: recall that the variance of a random variable Z with $\mathbb{E}[Z] = \mu$ is defined as $\text{Var}[Z] = \mathbb{E}[(Z - \mu)^2]$.

(Continued on page 2.)

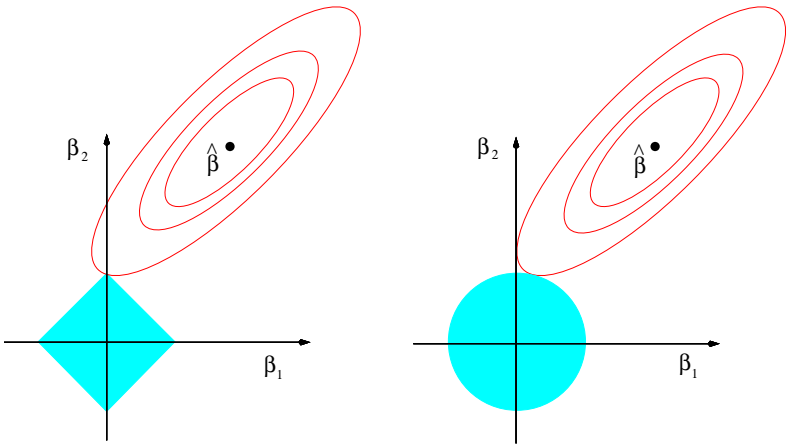

Figure 1: From Hastie et al. (2009). *The Elements of Statistical Learning.*

(b) Assuming that $\hat{f}(X)$ is a k-nearest-neighbor regression fit, the above expected prediction error expression takes the specific form of

$$\mathbb{E}[(Y - \hat{f}(x_0))^2 \,|\, x_0] = \sigma_\epsilon^2 + \left(f(x_0) - \frac{1}{k} \sum_{\ell=1}^{k} f(x_{(\ell)}) \right)^2 + \frac{\sigma_\epsilon^2}{k},$$

where $x_{(\ell)}$ denotes the ℓ:th nearest neighbor of x_0 in the training data. Describe the role of k in controlling the complexity of the model and the associated bias-variance tradeoff.

Problem 3 Linear Regression with Shrinkage

Assume a linear regression model $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$ with p input variables. Given some training data, we can define a shrinkage estimator of the model parameters as the solution to the following optimization problem:

$$\underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 \right\},$$

$$\text{subject to } R(\beta) \leq t,$$

where $R(\beta) = \sum_{j=1}^{p} \beta_j^2$ for ridge regression and $R(\beta) = \sum_{j=1}^{p} |\beta_j|$ for lasso.

(a) Describe briefly the general idea behind shrinkage methods, such as ridge regression and lasso, and also how t is connected to the bias-variance tradeoff for the fitted models.

(b) Consider Figure 1 which shows the contours (red ellipses) of the objective function around the least-squares solution $\hat{\beta}$ and the constraint regions of lasso and ridge regression (solid blue areas) for the above optimization problem in the case of two input variables. Based on Figure 1:

(Continued on page 3.)

(i) Explain the key difference between ridge regression and lasso in terms of the model they produce.

(ii) Explain how changing t would affect the constraint regions and the resulting parameter estimates.

Problem 4 Model Selection and Evaluation

(a) Consider the following strategy for performing a regression analysis in a case where there is a very large number of input variables $X_1, \ldots, X_p$:

1. Initial screening: Find a good subset of predictors by including the $q << p$ input variables that are most strongly correlated with the response variable.
2. K-fold cross-validation: For $k = 1, \ldots, K$, train a model for predicting the response given the q input variables (from Step 1) using all data except fold k and compute a fold-specific estimate of the test error based on the observations in fold k.
3. Average the fold-specific test errors obtained in Step 2 to obtain a cross-validation estimate of the test error.

Is the averaged test error obtained from the above procedure a good estimate of the expected prediction error of a new test observation from the same distribution? If not, what would you do differently to improve it?

(b) Let $\mathcal{T}$ denote a set of training data containing N joint observations of the input-output pair (X, Y) and let (X^0, Y^0) denote a new test data pair (all generated from the same $p(X, Y)$). Further let $\hat{f}(X)$ be a model fitted on the training data. Given some loss function $L(y, f(x))$, consider the following definitions of errors for the fitted model:

$$\text{(i)}\ \text{Err}_{\mathcal{T}} = \mathbb{E}_{X^0, Y^0}[L(Y^0, \hat{f}(X^0)) \,|\, \mathcal{T}]$$

$$\text{(ii)}\ \text{Err} = \mathbb{E}_{\mathcal{T}}\big[\mathbb{E}_{X^0, Y^0}[L(Y^0, \hat{f}(X^0)) \,|\, \mathcal{T}]\big]$$

$$\text{(iii)}\ \text{Err}_{\text{in}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{Y^0}[L(Y_i^0, \hat{f}(x_i)) \,|\, \mathcal{T}]$$

Explain what errors (i)–(iii) measure and what they are used for in the context of model selection and evaluation.

Problem 5 Bagged Trees and Random Forest

A regression and classification tree model can formally be expressed as

$$T(x; \Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j),$$

(Continued on page 4.)

where $\Theta = \{\gamma_j, R_j\}_{j=1}^{J}$ contains all the model parameters describing the tree, or its regions R_j, as well as the associated region-specific constants γ_j.

(a) By using bootstrap aggregation, or bagging, we can construct a so-called bagged tree estimate:

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x; \Theta_b).$$

How is the above estimate constructed and why does it in general improve the accuracy of a single-tree model?

(b) Let $Z_1, \ldots, Z_B$ be B identically distributed random variables with variance σ^2. If the variables are dependent, with a positive correlation ρ, one can show that:

$$\text{Var}\big[\frac{1}{B}\sum_{b=1}^{B} Z_b\big] = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

Explain why the bagged tree model is limited by the above result and how random forest can be considered an improved version of the bagged tree model.

(c) By letting $B \to \infty$, we obtain the limiting form of the random forest regression estimator:

$$\hat{f}_{\text{rf}}(x) = \mathbb{E}_{\Theta \mid \boldsymbol{Z}}[T(x; \Theta)],$$

where $\boldsymbol{Z}$ denotes the training data. Further, it can be shown that the total variance of a tree can be decomposed into a sum of two terms:

$$\text{Var}_{\Theta,\boldsymbol{Z}}\Big[T(x;\Theta)\Big] = \text{Var}_{\boldsymbol{Z}}\Big[\mathbb{E}_{\Theta \mid \boldsymbol{Z}}\big[T(x;\Theta)\big]\Big] + \mathbb{E}_{\boldsymbol{Z}}\Big[\text{Var}_{\Theta \mid \boldsymbol{Z}}\big[T(x;\Theta)\big]\Big].$$

What does the total variance and the two terms on the right-hand side of the above equation represent?

Problem 6 Boosting

Consider a two-class classification problem where the binary output variable is coded as $Y \in \{-1, 1\}$.

(a) For the considered problem, it can been shown that the AdaBoost algorithm in Figure 2 is equivalent to forward stagewise additive modelling under the exponential loss function:

$$L(y, f(x)) = e^{-yf(x)},$$

and, as a consequence of this, one can show that AdaBoost is seeking to estimate the population minimizer:

$$f^*(x) = \underset{f(x)}{\text{argmin}} \Big\{\mathbb{E}_{Y \mid x}\big[e^{-Yf(x)}\big]\Big\}.$$

Derive an expression for $f^*(x)$ and explain, based on the expression, why the classification rule of AdaBoost is reasonable.

(Continued on page 5.)

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 $G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$.

Figure 2: From Hastie et al. (2009). *The Elements of Statistical Learning.*

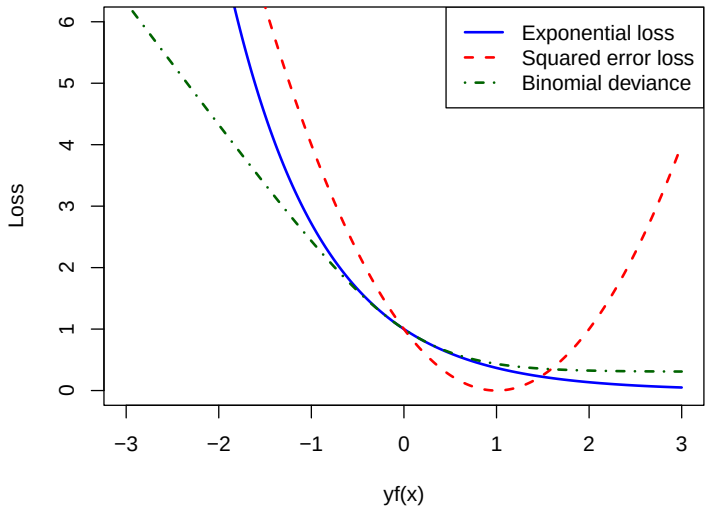

Figure 3: Different loss functions plotted against the margin.

(b) When classifying a $-1/1$ response using the classification rule $\text{sign}[f(x)]$, the margin $yf(x)$ plays a role analogous to the residuals in regression. Consider the exponential loss, squared error loss and binomial deviance, which are plotted in Figure 3 against the margin. Discuss the strengths and weaknesses of the considered loss functions in comparison to each other.

THE END - GOOD LUCK!