Theoretical task 7.

Recommendations: all solutions should be short, mathematically strict (unless qualitative explanation is needed), precise with respect to the stated question and clearly written. Solutions may be submitted in any readable format, including images.

1. Consider a regression problem where we need to predict value $f(x)$ of a single continuous variable x. Suppose we have M regression models that were trained on randomly generated data sets. The output $y_m(x)$ of each of the models can be written as

$$y_m(x) = f(x) + \epsilon_m(x) \tag{1}$$

where $\epsilon_m(x)$ is noise with Gaussian distribution $\mathcal{N}(\mu = 0, \sigma = 1)$. Consider the following bagging model:

$$y_{bagging}(x) = \frac{1}{M} \sum_{m=1}^{M} y_m(x) \tag{2}$$

Denote average prediction error for all models as:

$$E_{av} = \frac{1}{M} \sum_{m=1}^{M} E_x[(y_m(x) - f(x))^2] \tag{3}$$

Prediction error for the bagging model:

$$E_{bagging} = E_x[(y_{bagging}(x) - f(x))^2] \tag{4}$$

Prove the following expression:

$$E_{bagging} = \frac{1}{M} E_{av} \tag{5}$$

Suppose that noise values $\epsilon_m(x)$ for the models are uncorrelated.

2. Using Jensen's inequality show that $E_{bagging} \leq E_{av}$ for any convex error function $E(y)$, but not only squared error.

3. Using bias-variance decomposition for a regression problem from your lectures answer the following questions:

- What the smallest possible error for a regression model?
- How does bagging affect the decomposition? Why?

4. Recall bagging procedure. On lecture we have showed that if data sets (models) are i.i.d. then variance of averaged prediction is equal to $\frac{1}{B}\sigma^2$. However, in practice resulted data sets are not i.i.d. but only i.d. (identically distributed). Denote $\rho > 0$ as pairwise correlation between models in bagging ensemble. Prove that variance of bagging ensemble is equal to

$$Var[y_{bagging}] = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Hint: treat models as random variables with $E[x] = \mu$ and $Var[x] = \sigma^2$