Machine Learning B (2025)
Home Assignment 7

Carsten Jørgensen, student ID: skj730

Contents

1 XGBoost **2**

2 A simple version of Empirical Bernstein's inequality **3**

3 PAC-Bayes-Unexpected-Bernstein **6**

1 XGBoost

Question 1

The dataset contains 20,000 observations, so the training set consists of 16,000 rows, while the test set has 4,000 rows.

Question 2

The training set is further divided into a training and validation set, where 10% is used as a hold-out validation set. This split yields 1,600 validation observations. Using the parameters from the assignment (shown in Table 2), an XGBoost regression model was fitted to the remaining training data. The resulting training and validation RMSE values are shown in Figure 1, while the RMSE and (R^2) scores for the test set are reported in Table 1 under "Reference XGBoost“.

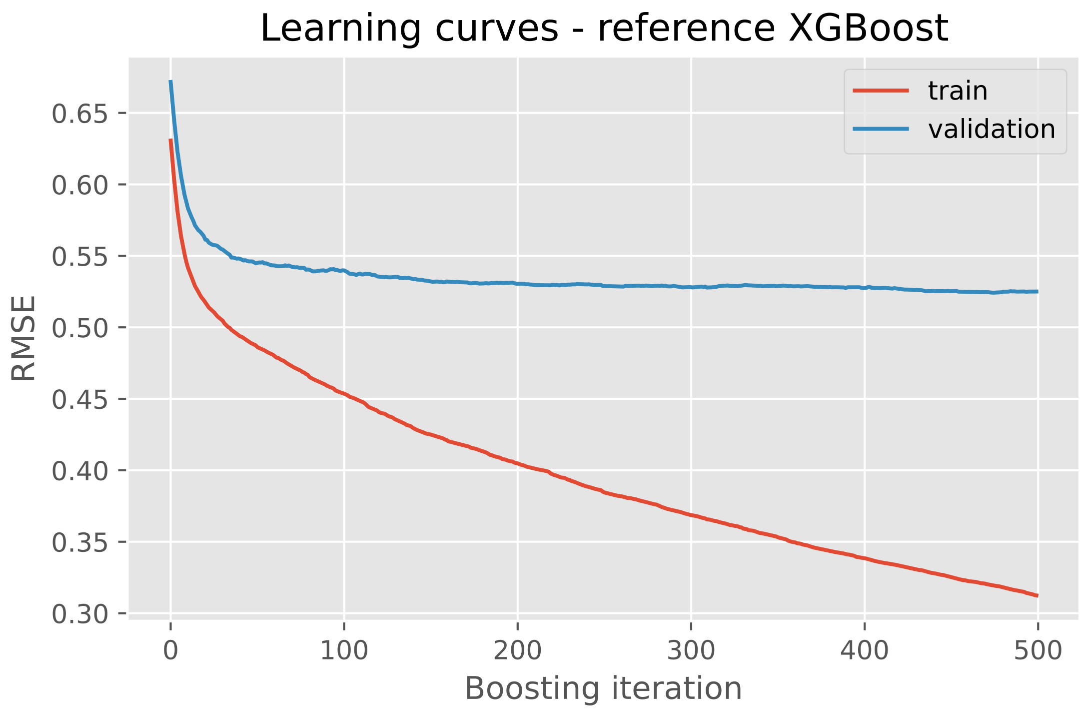

Figure 1: XGBoost Training vs. Validation RMSE

Table 1: Model comparison on the *unseen* test set (lower RMSE & higher R^2 are better)

Model	**RMSE**	**R^2**
Reference XGBoost	0.48234	0.36561
Tuned XGBoost	0.46141	0.41949
k-NN (k=5)	0.47985	0.37216

Table 2: XGBoost Model Parameters Comparison

Parameter	**Reference Model**	**Tuned Model**
`colsample_bytree`	0.5	0.6
`learning_rate`	0.1	0.01
`max_depth`	4	9
`n_estimators`	500	400
`reg_lambda`	1.0	1.5

Question 3

A grid search was used to find improved hyperparameters for the XGBoost model. The best parameter values are listed in Table 2, and Table 1 shows that the tuned XGBoost model performs slightly better than the reference model. It also outperforms a simple (k)-nearest-neighbors regression model.

Please refer to the code listing 1 to see essential parts of the code used to obtain these results.

2 A simple version of Empirical Bernstein's inequality

The quadratic-difference identity

Let X and X' be two independent identically distributed random variables. Prove that $\mathbb{E}\big[(X-X')^2\big] = 2\,\mathbb{V}[X]$.

Proof.

$$\begin{aligned}
\mathbb{E}\left[(X-X')^2\right] &= \mathbb{E}\left[X^2\right] + \mathbb{E}\left[X'^2\right] - 2\mathbb{E}\left[XX'\right] \\
&= 2\mathbb{E}\left[X^2\right] - 2\big(\mathbb{E}\left[X\right]\big)^2 \qquad \text{(independence of } X \text{ and } X'\text{)} \\
&= 2\,\mathbb{V}[X].
\end{aligned}$$

□

High-probability upper bound on the variance

Let $X_1, \ldots, X_n$ be independent identically distributed random variables taking values in the $[0,1]$ interval, and assume that n is even. Let $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n/2}(X_{2i} - X_{2i-1})^2$ and let $\nu = \mathbb{V}[X_1]$. Then we have:

$$P\left(\nu \geq \hat{\nu}_n + \sqrt{\frac{\ln\frac{1}{\delta}}{n}}\right) \leq \delta.$$

Proof. Let us define $Y_i = (X_{2i} - X_{2i-1})^2$ for $i = 1, \dots, n/2$. From question 1, we know that $\mathbb{E}\left[(X_{2i} - X_{2i-1})^2\right] = 2\,\mathbb{V}[X_1] = 2\nu$. Therefore:

$$\mathbb{E}\left[\hat{\nu}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n/2} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n/2}\mathbb{E}\left[Y_i\right] = \frac{1}{n}\sum_{i=1}^{n/2}(2\nu) = \frac{1}{n}\cdot\frac{n}{2}\cdot 2\nu = \nu$$

So $\hat{\nu}_n$ is an unbiased estimator of ν.

Since each $X_j \in [0,1]$, we have $|X_{2i} - X_{2i-1}| \leq 1$, so $Y_i = (X_{2i} - X_{2i-1})^2 \in [0,1]$. The variables $Y_1, \dots, Y_{n/2}$ are independent (since they involve disjoint pairs of the original variables), each bounded in $[0,1]$, with $\mathbb{E}\left[Y_i\right] = 2\nu$.
We can write:

$$\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n/2} Y_i = \frac{1}{2}\cdot\frac{1}{n/2}\sum_{i=1}^{n/2} Y_i$$

Let $W = \frac{1}{n/2}\sum_{i=1}^{n/2} Y_i$. Then $\hat{\nu}_n = \frac{1}{2}W$ and $\mathbb{E}\left[W\right] = 2\nu$. Applying Hoeffding's inequality to the empirical average W:

$$P[W \leq 2\nu - \varepsilon] \leq e^{-2(n/2)\varepsilon^2} = e^{-n\varepsilon^2}$$

Since $\hat{\nu}_n = \frac{1}{2}W$:

$$P\left[\hat{\nu}_n \leq \nu - \frac{\varepsilon}{2}\right] \leq e^{-n\varepsilon^2}$$

Define $e^{-n\varepsilon^2} = \delta$ gives $\varepsilon = \sqrt{\frac{\ln(1/\delta)}{n}}$.
Therefore:

$$P\left(\hat{\nu}_n \leq \nu - \frac{1}{2}\sqrt{\frac{\ln\frac{1}{\delta}}{n}}\right) \leq \delta$$

and after rearranging:

$$P\left(\nu \geq \hat{\nu}_n + \frac{1}{2}\sqrt{\frac{\ln\frac{1}{\delta}}{n}}\right) \leq \delta$$

Since $\frac{1}{2}\sqrt{x} < \sqrt{x}$, we have:

$$\left\{\nu \geq \hat{\nu}_n + \sqrt{\frac{\ln\frac{1}{\delta}}{n}}\right\} \subseteq \left\{\nu \geq \hat{\nu}_n + \frac{1}{2}\sqrt{\frac{\ln\frac{1}{\delta}}{n}}\right\}$$

Therefore:

$$P\left(\nu \geq \hat{\nu}_n + \sqrt{\frac{\ln\frac{1}{\delta}}{n}}\right) \leq P\left(\nu \geq \hat{\nu}_n + \frac{1}{2}\sqrt{\frac{\ln\frac{1}{\delta})}{n}}\right) \leq \delta.$$

$\square$

A weaker empirical Bernstein inequality

Let $X_1, \dots, X_n$, n, ν, and $\hat{\nu_n}$ be as before, and let $\mu = \mathbb{E}[X_1]$. Then we have:

$$\mathbb{P}\left(\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{2\hat{\nu}_n \ln\frac{2}{\delta}}{n}} + \sqrt{2}\left(\frac{\ln\frac{2}{\delta}}{n}\right)^{\frac{3}{4}} + \frac{\ln\frac{2}{\delta}}{3n}\right) \leq \delta$$

Proof. Let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and define the following events:

$$A = \left\{\nu \geq \hat{\nu}_n + \sqrt{\frac{2\hat{\nu}_n \ln\frac{2}{\delta}}{n}} + \sqrt{2}\left(\frac{\ln\frac{2}{\delta}}{n}\right)^{\frac{3}{4}} + \frac{\ln\frac{2}{\delta}}{3n}\right\}$$

$$B = \left\{\nu \leq \hat{\nu}_n + \sqrt{\frac{\ln\frac{2}{\delta}}{n}}\right\}$$

For all sets A and B we have $\mathbb{P}(A) \leq \mathbb{P}(A|B) + \mathbb{P}(\overline{B})$ where $\overline{B}$ is the complement event to B.

Bound $\mathbb{P}(\overline{B})$:

From question 2 with confidence parameter $\delta/2$:

$$\mathbb{P}(\overline{B}) = \mathbb{P}\left(\nu > \hat{\nu}_n + \sqrt{\frac{\ln\frac{2}{\delta}}{n}}\right) \leq \frac{\delta}{2}$$

Bound $\mathbb{P}(A|B)$:

From Bernstein's inequality (Theorem 2.38 in [Sel25]) with confidence $\delta/2$:

$$\mathbb{P}\left(\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{2\nu \ln\frac{2}{\delta}}{n}} + \frac{\ln\frac{2}{\delta}}{3n}\right) \leq \frac{\delta}{2}$$

Given that event B occurs, we have $\nu \leq \hat{\nu}_n + \sqrt{\frac{\ln(2/\delta)}{n}}$, so:

$$\sqrt{\frac{2\nu \ln\frac{2}{\delta}}{n}} \leq \sqrt{\frac{2\left(\hat{\nu}_n + \sqrt{\frac{\ln\frac{2}{\delta}}{n}}\right)\ln\frac{2}{\delta}}{n}}$$

Using the hint $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$:

$$\sqrt{\frac{2\left(\hat{\nu}_n + \sqrt{\frac{\ln\frac{2}{\delta}}{n}}\right)\ln\frac{2}{\delta}}{n}} \leq \sqrt{\frac{2\hat{\nu}_n \ln\frac{2}{\delta}}{n}} + \sqrt{\frac{2\sqrt{\frac{\ln\frac{2}{\delta}}{n}}\ln\frac{2}{\delta}}{n}}$$

The second term can be rewritten as:

$$\sqrt{\frac{2\sqrt{\frac{\ln\frac{2}{\delta}}{n}}\ln\frac{2}{\delta}}{n}} = \sqrt{2\cdot\frac{(\ln\frac{2}{\delta})^{3/2}}{n^{3/2}}} = \sqrt{2}\cdot\left(\frac{\ln\frac{2}{\delta}}{n}\right)^{3/4}$$

Therefore, given that B occurs:

$$\mathbb{P}\left(\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{2\hat{\nu}_n \ln\frac{2}{\delta}}{n}} + \sqrt{2}\left(\frac{\ln\frac{2}{\delta}}{n}\right)^{3/4} + \frac{\ln\frac{2}{\delta}}{3n} \,\middle|\, B\right) \leq \frac{\delta}{2}$$

Combine the bounds

$$\mathbb{P}(A) \leq \mathbb{P}(A|B) + \mathbb{P}(\bar{B}) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

Therefore:

$$\mathbb{P}\left(\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{2\hat{\nu}_n \ln\frac{2}{\delta}}{n}} + \sqrt{2}\left(\frac{\ln\frac{2}{\delta}}{n}\right)^{\frac{3}{4}} + \frac{\ln\frac{2}{\delta}}{3n}\right) \leq \delta.$$

□

3 PAC-Bayes-Unexpected-Bernstein

Proposition 1

For any random variable $Z \leq 1$ and $\lambda \in [0, \frac{1}{2}]$ we have:

$$\mathbb{E}\left[e^{-\lambda Z - \lambda^2 Z^2}\right] \leq e^{-\lambda \mathbb{E}[Z]}.$$

Proof. Since $Z \leq 1$ and $\lambda \leq \frac{1}{2}$ we have $\lambda Z \leq \frac{1}{2}$, so $-\lambda Z \geq -\frac{1}{2}$. Using the hint that $z - z^2 \leq \ln(1+z)$ for $z \geq -\frac{1}{2}$ with $z = -\lambda Z$:

$$-\lambda Z - (-\lambda Z)^2 = -\lambda Z - \lambda^2 Z^2 \leq \ln(1 - \lambda Z)$$

Taking exponentials:

$$e^{-\lambda Z - \lambda^2 Z^2} \leq e^{\ln(1-\lambda Z)} = 1 - \lambda Z$$

Taking expectations:

$$\mathbb{E}\left[e^{-\lambda Z - \lambda^2 Z^2}\right] \leq \mathbb{E}\left[1 - \lambda Z\right] = 1 - \lambda \mathbb{E}\left[Z\right]$$

Now we use the second hint that $1 + z \leq e^z$ with $z = -\lambda \mathbb{E}\left[Z\right]$:

$$1 - \lambda \mathbb{E}\left[Z\right] \leq e^{-\lambda \mathbb{E}[Z]}$$

Hence:

$$\mathbb{E}\left[e^{-\lambda Z-\lambda^2 Z^2}\right] \leq e^{-\lambda \mathbb{E}[Z]}.$$

The assumption $\lambda \in [0, \frac{1}{2}]$ ensures $-\lambda Z \geq -\frac{1}{2}$ for the first inequality, and $Z \leq 1$ is used to bound λZ.

□

Proposition 2

For $Z \leq 1$ and $\lambda \in [0, \frac{1}{2}]$:

$$\mathbb{E}\left[e^{\lambda(\mathbb{E}[Z]-Z)-\lambda^2 Z^2}\right] \leq 1$$

Proof. We can rewrite the left side as:

$$\mathbb{E}\left[e^{\lambda(\mathbb{E}[Z]-Z)-\lambda^2 Z^2}\right] = \mathbb{E}\left[e^{\lambda\mathbb{E}[Z]-\lambda Z-\lambda^2 Z^2}\right] = e^{\lambda\mathbb{E}[Z]}\mathbb{E}\left[e^{-\lambda Z-\lambda^2 Z^2}\right]$$

From Exercise 1, we have $\mathbb{E}\left[e^{-\lambda Z-\lambda^2 Z^2}\right] \leq e^{-\lambda\mathbb{E}[Z]}$. Hence:

$$\mathbb{E}\left[e^{\lambda(\mathbb{E}[Z]-Z)-\lambda^2 Z^2}\right] = e^{\lambda\mathbb{E}[Z]}\mathbb{E}\left[e^{-\lambda Z-\lambda^2 Z^2}\right] \leq e^{\lambda\mathbb{E}[Z]}e^{-\lambda\mathbb{E}[Z]} = 1$$

□

Proposition 3

For independent random variables $Z_1, \ldots, Z_n$ upper bounded by 1 and $\lambda \in [0, \frac{1}{2}]$:

$$\mathbb{E}\left[e^{\lambda\sum_{i=1}^n(\mathbb{E}[Z_i]-Z_i)-\lambda^2\sum_{i=1}^n Z_i^2}\right] \leq 1.$$

Proof. Using independence:

$$\mathbb{E}\left[e^{\lambda\sum_{i=1}^n(\mathbb{E}[Z_i]-Z_i)-\lambda^2\sum_{i=1}^n Z_i^2}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda(\mathbb{E}[Z_i]-Z_i)-\lambda^2 Z_i^2}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(\mathbb{E}[Z_i]-Z_i)-\lambda^2 Z_i^2}\right]$$

From Proposition 2, each factor satisfies $\mathbb{E}\left[e^{\lambda(\mathbb{E}[Z_i]-Z_i)-\lambda^2 Z_i^2}\right] \leq 1$.
Therefore:

$$\prod_{i=1}^n \mathbb{E}\left[e^{\lambda(\mathbb{E}[Z_i]-Z_i)-\lambda^2 Z_i^2}\right] \leq 1$$

□

Proposition 4

For independent $Z_1, \ldots, Z_n$ upper bounded by 1 and $\lambda \in (0, \frac{1}{2}]$:

$$P\left(E\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] \geq \frac{1}{n}\sum_{i=1}^{n} Z_i + \frac{\lambda}{n}\sum_{i=1}^{n} Z_i^2 + \frac{\ln\frac{1}{\delta}}{\lambda n}\right) \leq \delta$$

Proof. From Proposition 3: $\mathbb{E}\left[e^{\lambda\sum_{i=1}^{n}(E[Z_i]-Z_i)-\lambda^2\sum_{i=1}^{n}Z_i^2}\right] \leq 1$ Using Markov's inequality:

$$\mathbb{P}\left(\lambda\sum_{i=1}^{n}(\mathbb{E}[Z_i] - Z_i) - \lambda^2\sum_{i=1}^{n} Z_i^2 \geq \ln\tfrac{1}{\delta}\right) \leq \delta$$

and rearranging:

$$\mathbb{P}\left(\mathbb{E}\left[\sum_{i=1}^{n} Z_i\right] \geq \sum_{i=1}^{n} Z_i + \lambda\sum_{i=1}^{n} Z_i^2 + \frac{\ln\frac{1}{\delta}}{\lambda}\right) \leq \delta$$

Now dividing by n gives the desired result. □

Proposition 5

For grid $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ with $\lambda_i \in (0, \frac{1}{2}]$:

$$\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] \geq \frac{1}{n}\sum_{i=1}^{n} Z_i + \min_{\lambda\in\Lambda}\left(\frac{\lambda}{n}\sum_{i=1}^{n} Z_i^2 + \frac{\ln\frac{k}{\delta}}{\lambda n}\right)\right) \leq \delta$$

Proof. From Part 4, for each $\lambda \in \Lambda$ with confidence δ/k:

$$\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] \geq \frac{1}{n}\sum_{i=1}^{n} Z_i + \frac{\lambda}{n}\sum_{i=1}^{n} Z_i^2 + \frac{\ln\frac{k}{\delta}}{\lambda n}\right) \leq \frac{\delta}{k}$$

The event that the inequality holds for the minimum over Λ is contained in the union of events where it holds for each individual λ. By the union bound:

$$\mathbb{P}\left(\bigcup_{\lambda\in\Lambda}\left\{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] \geq \frac{1}{n}\sum_{i=1}^{n} Z_i + \frac{\lambda}{n}\sum_{i=1}^{n} Z_i^2 + \frac{\ln\frac{k}{\delta}}{\lambda n}\right\}\right) \leq \sum_{\lambda\in\Lambda}\frac{\delta}{k} = \delta$$

□

Experimental evaluation 6

This part requires implementation and empirical comparison. The Unexpected Bernstein bound should be tighter when the second moment $\hat{v}_n$ is small relative to the first moment $\hat{p}_n$, as it can exploit the variance structure that the kl inequality cannot capture. We see in Figure 2 that the Unexpected Bernstein is decreasing when the second moment decreases.
In short: the kl inequality is variance-blind and gives a single universal radius, whereas the Unexpected-Bernstein inequality adapts to the empirical variance and is substantially tighter whenever the data are concentrated.

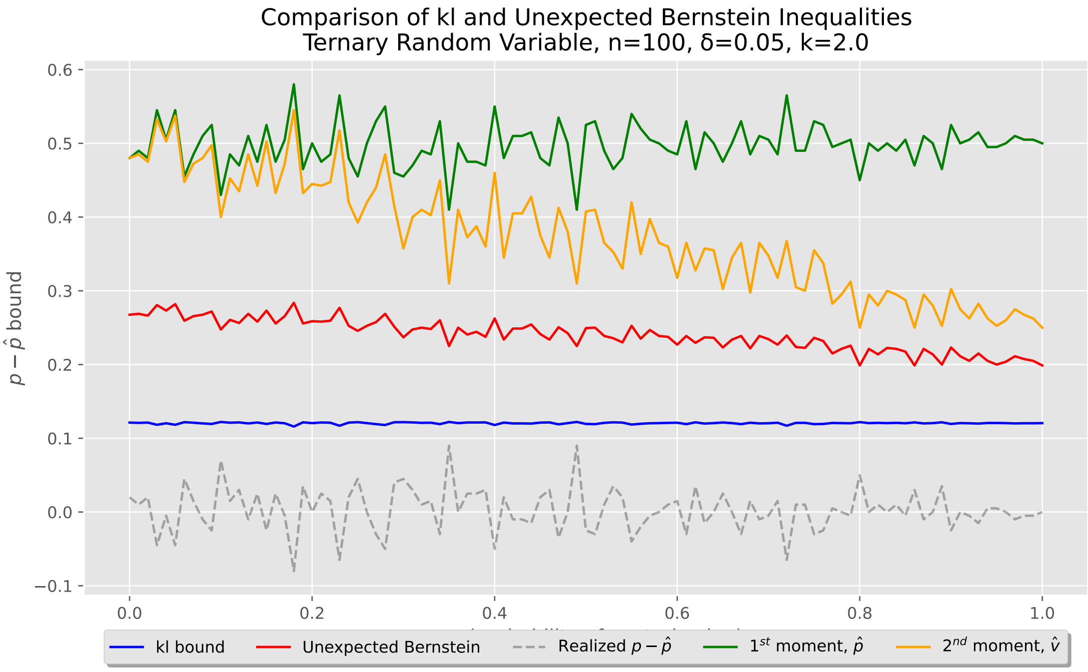

Figure 2: Unexpected Bernstein vs kl bound

Proposition 7

For sample i.i.d. sample S, prediction rule h, and loss function ℓ upper bounded by 1:

$$\mathbb{E}\left[e^{n(\lambda(L(h)-\hat{L}(h,S))-\lambda^2\hat{V}(h,S))}\right] \leq 1$$

where $\hat{V}(h,S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)^2$ and $\lambda \in [0, \frac{1}{2}]$.

Proof. If we define $Z_i = \ell(h(X_i), Y_i)$, we have:

- $\mathbb{E}\left[Z_i\right] = L(h)$
- $\frac{1}{n}\sum_{i=1}^{n} Z_i = \hat{L}(h,S)$
- $\frac{1}{n}\sum_{i=1}^{n} Z_i^2 = \hat{V}(h,S)$

Using Proposition 3:

$$\mathbb{E}\left[e^{\lambda\sum_{i=1}^{n}(\mathbb{E}[Z_i]-Z_i)-\lambda^2\sum_{i=1}^{n} Z_i^2}\right] \leq 1$$

Substituting and isolating n gives the desired result:

$$\mathbb{E}\left[e^{n\lambda(L(h)-\hat{L}(h,S))-n\lambda^2\hat{V}(h,S)}\right] = \mathbb{E}\left[e^{n(\lambda(L(h)-\hat{L}(h,S))-\lambda^2\hat{V}(h,S))}\right] \leq 1$$

$\square$

Proposition 8

For hypothesis set $\mathcal{H}$, prior π independent of S, and $\lambda \in (0, \frac{1}{2}]$:

$$\mathbb{P}\left(\exists\rho : \mathbb{E}_\rho[L(h)] \geq \mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \lambda\mathbb{E}_\rho\left[\hat{V}(h,S)\right] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{n\lambda}\right) \leq \delta$$

Proof. Let

$$f(h,S) = n\Big(\lambda\big(L(h) - \widehat{L}(h,S)\big) - \lambda^2\widehat{V}(h,S)\Big).$$

and apply the PAC-Bayes lemma (Theorem 3.27 in [Sel25]) with $f(h,S)$ to get:

$$\mathbb{E}_\rho[f(h,S)] \leq \mathrm{KL}(\rho\|\pi) + \ln\mathbb{E}_\pi\left[e^{f(h,S)}\right] \tag{1}$$

Taking expectation over S in the result from Proposition 7 gives $\mathbb{E}_{S,\pi}[e^{f(h,S)}] = \mathbb{E}_{\pi,S}[e^{f(h,S)}] \leq 1$[1]. Hence for every fixed sample S

$$\mathbb{E}_\pi\left[e^{f(h,S)}\right] \leq \frac{1}{\delta} \quad \text{with probability } 1-\delta$$

by Markov's inequality (same argument as in Lemma 3.28). Thus with probability at least $1-\delta$ over S, for all ρ

$$\mathbb{E}_\rho[f(h,S)] \leq \mathrm{KL}(\rho\|\pi) + \ln\tfrac{1}{\delta}. \tag{2}$$

Divide (2) by $(n\lambda > 0)$ and use the definition of $f(h,S)$:

$$\mathbb{E}_\rho[L(h)] \leq \mathbb{E}_\rho\big[\hat{L}(h,S)\big] + \lambda\mathbb{E}_\rho\big[\hat{V}(h,S)\big] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{n\lambda}.$$

Because we obtained it with probability at least $1-\delta$ and then took the union bound over all ρ (again as in Lemma 3.28), the probability of violation is at most δ, which proves the Proposition. □

Proposition 9

For grid $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ with $\lambda_i \in (0, \frac{1}{2}]$:

$$\mathbb{P}\left(\exists\rho : \mathbb{E}_\rho\Big[L(h)\Big] \geq \mathbb{E}_\rho\Big[\hat{L}(h,S)\Big] + \min_{\lambda\in\Lambda}\left(\lambda\mathbb{E}_\rho\Big[\hat{V}(h,S)\Big] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{k}{\delta}}{n\lambda}\right)\right) \leq \delta$$

Proof. From Proposition 8, for each $\lambda \in \Lambda$ with confidence δ/k:

$$\mathbb{P}\left(\exists\rho : \mathbb{E}_\rho\Big[L(h)\Big] \geq \mathbb{E}_\rho\Big[\hat{L}(h,S)\Big] + \lambda\mathbb{E}_\rho\Big[\hat{V}(h,S)\Big] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{k}{\delta}}{n\lambda}\right) \leq \frac{\delta}{k}$$

Using the same arguments as in Proposition 5. The event that the inequality holds for the minimum over Λ is contained in the union of events where it holds for each individual λ, so taking the union bound over all $\lambda \in \Lambda$ gives the desired result. □

[1] Here I assume we can use Fubini's theorem to switch order of integration for S and π

Appendix A

Code listing for essential parts of the XGBoost algorithm

```
eval_metric = "rmse"

xgb_ref = XGBRegressor(
    objective="reg:squarederror",
    # parameters specified by the exercise
    colsample_bytree=0.5,
    learning_rate=0.1,
    max_depth=4,
    reg_lambda=1.0,
    n_estimators=500,
    eval_metric=eval_metric,
    # nice-to-have settings
    random_state=RANDOM_STATE,
    verbosity=0,
)

eval_set = [(X_tr, y_tr), (X_val, y_val)]

xgb_ref.fit(X_tr, y_tr, eval_set=eval_set, verbose=False)

# --- Plot RMSE vs. boosting iteration ------------------------
results = xgb_ref.evals_result()
rmse_tr = results["validation_0"][eval_metric]
rmse_va = results["validation_1"][eval_metric]

# --- Evaluate on the *held-out* test set ---------------------
y_pred_ref = xgb_ref.predict(X_test)
rmse_ref = root_mean_squared_error(y_test, y_pred_ref)
r2_ref = r2_score(y_test, y_pred_ref)

# -------------------------------------------------------------
# Step 3.  Hyper-parameter grid search (3-fold CV) plus KNN = 5
# -------------------------------------------------------------
param_grid = {
    "colsample_bytree": [0.5, 0.6],
    "learning_rate": [0.001, 0.01, 0.02],
    "max_depth": [8, 9, 10],
    "n_estimators": [400, 500],
    "reg_lambda": [1.5, 1.6],
}

xgb_base = XGBRegressor(
    objective="reg:squarederror", random_state=RANDOM_STATE,
   verbosity=0
)

grid = GridSearchCV(estimator=xgb_base, param_grid=param_grid,
    cv=3, scoring="neg_root_mean_squared_error", n_jobs=-1,
    verbose=1)
```

```python
grid.fit(X_train, y_train)  # ONLY on *training* part

best_params = grid.best_params_
best_rmse = -grid.best_score_  # negate again

# ---- Train final XGBoost with the discovered configuration --
xgb_best = XGBRegressor(
    objective="reg:squarederror", random_state=RANDOM_STATE,
   verbosity=0, **best_params
)
xgb_best.fit(X_train, y_train)

y_pred_best = xgb_best.predict(X_test)
rmse_best = root_mean_squared_error(y_test, y_pred_best)
r2_best = r2_score(y_test, y_pred_best)

# ---- Baseline: 5-nearest-neighbours regressor ------------
knn = KNeighborsRegressor(n_neighbors=5)
knn.fit(X_train, y_train)
y_pred_knn = knn.predict(X_test)
rmse_knn = root_mean_squared_error(y_test, y_pred_knn)
r2_knn = r2_score(y_test, y_pred_knn)
```

Listing 1: XGBoost

References

[Sel25] Yevgeny Seldin. *Machine Learning The Science of Selection under Uncertainty*. June 2, 2025. URL: https://sites.google.com/site/yevgenyseldin/teaching.