Machine Learning for EDS

Tutorial week 3 : Solutions

2023/2024

Important: this file is meant for students of the course Machine Learning for EDS (2023/2024) and is not allowed to be distributed to others.

Generically in the following exercises, consider $\mathcal{X}$ to be a measurable space of features, $\mathcal{Y}$ a measurable space of outputs, P a distribution over $\mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim P$, $\eta(X) = \mathbb{E}[Y|X]$ and $\mathcal{F}$ is the set of all predictors (that is, measurable functions) from $\mathcal{X}$ into $\mathcal{Y}$.

Problem 1

1. The Bayes predictor is equal to $\eta(X) = \mathbb{E}[Y|X] = \mathbb{E}[aX_1 + bX_1X_2 + \varepsilon|X] = aX_1 + bX_1X_2 + \mathbb{E}[\varepsilon] = aX_1 + bX_1X_2$, using that ε has mean zero and is independent of X. Hence, Bayes risk is equal to $\mathbb{E}[(Y - \eta(X))^2] = \mathbb{E}[(aX_1 + bX_1X_2 + \varepsilon - aX_1 - bX_1X_2)^2] = \mathbb{E}[\varepsilon^2] = \sigma^2$.

2. By definition, the empirical risk of some predictor f is equal to $\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2$. So for $f_d \in S$, we have that $\widehat{\mathcal{R}}_n(f_d) = \frac{1}{n}\sum_{i=1}^n (y_i - dx_{i1})^2$.

3. The function $\widehat{\mathcal{R}}_n(f_d)$ is clearly strictly convex in d, as it is a sum of strictly convex functions. Hence, the critical point of the function is its global minimum. Taking the derivative of the empirical risk with respect to d, leads to $-2\frac{1}{n}\sum_{i=1}^n x_{i1}(y_i - dx_{i1})$. Setting this derivative to zero, and solving for d leads to:

$$-2\frac{1}{n}\sum_{i=1}^n x_{i1}(y_i - \hat{d}x_{i1}) = 0 \iff \sum_{i=1}^n x_{i1}y_i = \hat{d}\sum_{i=1}^n x_{i1}^2 \iff \hat{d} = \frac{\sum_{i=1}^n x_{i1}y_i}{\sum_{i=1}^n x_{i1}^2}.$$

4. Using the definition of expected risk, we have that

$$\begin{aligned}
\mathcal{R}_P(f_{\hat{d}}) &= \mathbb{E}[(Y - f_{\hat{d}}(X))^2|D_n] \\
&= \mathbb{E}[(Y - \hat{d}X_1)^2|D_n] \\
&= \mathbb{E}[(aX_1 + bX_1X_2 + \varepsilon - \hat{d}X_1)^2|D_n] \\
&= \mathbb{E}[X_1^2(a - \hat{d} + bX_2)^2|D_n] + \mathbb{E}[\varepsilon^2] + 2\mathbb{E}[\varepsilon X_1(a - \hat{d} + bX_2)|D_n] \\
&= \mathbb{E}[X_1^2]\mathbb{E}[(a - \hat{d} + bX_2)^2|D_n] + \mathbb{E}[\varepsilon^2] + 2\mathbb{E}[\varepsilon]\mathbb{E}[X_1(a - \hat{d} + bX_2)|D_n] \\
&= \mathbb{E}[X_1^2]\mathbb{E}[(a - \hat{d} + bX_2)^2|D_n] + \mathbb{E}[\varepsilon^2]\,,
\end{aligned}$$

where in the fourth equality we use that ε is independent of the sample D_n and in the fifth equality we use that X_1 is assumed to be independent of X_2 and ε is assumed to be independent of X. Finally, we use that $\mathbb{E}[\varepsilon] = 0$ and that $\mathbb{E}[X_i^2] < \infty$ for $i = 1, 2$ by assumption.

Because Bayes risk is equal to $\mathcal{R}_P^* = \mathbb{E}[\varepsilon^2]$, it follows immediately that the excess risk of $f_{\hat{d}}$ is equal to $\ell(f^*, f_{\hat{d}}) = \mathcal{R}_P(f_{\hat{d}}) - \mathcal{R}_P^* = \mathbb{E}[X_1^2]\mathbb{E}[(a - \hat{d} + bX_2)^2|D_n]$.

Under the assumption that $\mathbb{E}[X_2] = 0$, excess risk can be rewritten as $\ell(f^*, f_{\hat{d}}) = \mathbb{E}[X_1^2]\left((a - \hat{d})^2 + b^2\mathbb{E}[X_2^2]\right)$, so it is immediately clear that this quantity can only be equal to zero if either (i) $\hat{d} = a$ and $b = 0$ or (ii) $\mathbb{E}[X_1^2] = 0$. This is sensible, because in case (i) we have $Y = aX_1 + \varepsilon$, so the Bayes predictor becomes $\eta(X) = aX_1$ and the model S contains this predictor, because $f_a(X) = aX_1$ is an element of the model S. In case (ii), it follows that X_1 equals zero with probability one, which implies that $f_d(X) = dX_1$ will also be zero with probability one, regardless of the value of d. Also, in this case effectively $Y = \varepsilon$ almost surely, which implies that each predictor in the model S will have risk $\mathbb{E}[\varepsilon^2]$, i.e. Bayes risk.

5. $\hat{d}$ is a random variable because the empirical risk $\hat{\mathcal{R}}_n(f)$ and hence the empirical risk minimizer $\hat{d}$, depend on the sample D_n which is random. It follows that also $\mathcal{R}_P(f_{\hat{d}})$ is random, as it is an expectation of a function of $\hat{d}$, conditional on the random sample D_n.

6. (1) Because the distribution P of the features and labels (X, Y) is in practice unknown, we cannot calculate the precise value of the risk. (2) Because in most cases we do not have a closed form expression of the empirical risk minimizer $\hat{f}$. For example if the model is non-linear in the features, then the solution of the minimization of the empirical risk is usually not available in closed form and we have to use numerical minimization to find the minimizer of empirical risk.

Problem 2

1. We have shown in the lectures that the expression of Bayes risk in this framework is given by:

$$\mathcal{R}_p^* = \inf_{f \in \mathcal{F}} \mathcal{R}_p^c(f) = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}]$$

We also showed that a Bayes classifier/target function is $f^* : x \mapsto \mathbb{1}_{\eta(x) > 1/2}$, and that this is the only Bayes classifier except for in the event $\{\eta(X) = 1/2\}$. Finally, we have shown that, conditionally on a sample, the excess risk of a sample-dependent predictor f writes

$$\ell(f^*, f(D_n)) = \mathbb{E}\Big[|2\eta(X) - 1|\mathbb{1}_{f(D_n;X) \neq f^*(X)}\Big|D_n\Big],$$

$$= 2\mathbb{E}\Big[|\eta(X) - 1/2|\mathbb{1}_{f(D_n;X)\neq f^*(X)}\Big|D_n\Big], \tag{1}$$

where f^* is again the Bayes classifier.

2. (a) Because $\eta(X) \in \{0,1\}$ where $\eta(X) = \mathbb{E}(Y|X) = \mathbb{P}(Y=1|X)$, we know that

$$\eta(x) = \mathbb{E}(Y|X=x) = \mathbb{P}(Y=1|X=x)$$

is either zero or one depending on x. In other words, conditional on X,

$$\eta(X) = 1 \iff Y = 1 \quad \text{with probability one}$$

and

$$\eta(X) = 0 \iff Y = 0 \quad \text{with probability one.}$$

Recall, $f^* : x \mapsto \mathbb{1}_{\eta(x)>1/2}$. Hence:

$$\begin{aligned} f^*(X) &= \mathbb{1}_{\eta(X)>1/2} \\ &= \left(\mathbb{1}_{\eta(X)=1} + \mathbb{1}_{\eta(X)=0}\right)\mathbb{1}_{\eta(X)>1/2} \\ &= \mathbb{1}_{\eta(X)=1} = \mathbb{1}_{Y=1} = Y \end{aligned}$$

almost surely, where the second equality is valid because $\eta(X) \in \{0,1\}$ and the fourth equality holds because $\eta(X) = 1 \iff Y = 1$ a.s. and $\eta(X) \neq 1 \iff Y = 0$ a.s. Or in words: if $\eta(x)$ equals 1, then $\eta(x) = \mathbb{P}(Y=1|X=x) = 1 > 1/2$, so $f^*(x) = 1 = Y$ almost surely. If $\eta(x) = \mathbb{P}(Y=1|X=x) = 0$, then $\eta(x) = 0 \leq 1/2$, so then $f^*(x) = 0 = Y$ almost surely. In conclusion: $f^*(X) = Y$ almost surely.

The Bayes risk is equal to $\mathcal{R}_p^* = \mathbb{E}[\min\{\eta(X), 1-\eta(X)\}] = 0$, since $\eta(X) \in \{0,1\}$. The interpretation of the zero error assumption is that the value of Y is known given X, so after 'controlling' for X there is no randomness left in Y. This is a restrictive assumption in practice, because it essentially assumes there is a deterministic relationship between X and Y, which is often not the case in classification problems. On the other hand, there are examples where this assumption is realistic.

(b) The plug-in classifier associated to $\hat{\eta}$ is:

$$\hat{f}_{\hat{\eta}}(D_n; x) = \mathbb{1}_{\hat{\eta}(D_n;x)>1/2}\,.$$

(c) If $\hat{f}_{\hat{\eta}}(D_n; X) \neq f^*(X)$ then either $\hat{\eta}(D_n; x) > 1/2$ and $\eta(x) = 0 \leq 1/2$ or $\hat{\eta}(D_n; x) \leq 1/2$ and $\eta(x) = 1 > 1/2$. So it follows automatically that if $\hat{f}_{\hat{\eta}}(D_n; X) \neq f^*(X)$, then either

$$\hat{\eta}(D_n; X) \leq \frac{1}{2} < \eta(X) \qquad \text{or} \qquad \eta(X) \leq \frac{1}{2} < \hat{\eta}(D_n; X).$$

Note: without the zero-error assumption the statement we have to prove would also hold.

(d) By the previous point we know that if $\hat{f}_{\hat{\eta}}(D_n; X) \neq f^*(X)$ then $\eta(X)$ and $\hat{\eta}(D_n; X)$ are further apart than $\eta(X)$ and $1/2$, so that $|\eta(X) - \hat{\eta}(D_n; X)| \geq |\eta(X) - \frac{1}{2}|$. The results now follows directly from this observation:

$$2\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X) \neq f^*(X)} \leq 2|\hat{\eta}(D_n; X) - \eta(X)| \mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X) \neq f^*(X)} \tag{2}$$

(e) Using the expression of excess risk we gave in question 2.1, see (1), and plugging in the result in (2) gives:

$$\begin{aligned}
\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) &= 2\mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X) \neq f^*(X)} \Big| D_n\right] \\
&\leq 2\mathbb{E}\Big[|\hat{\eta}(D_n; X) - \eta(X)| \mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X) \neq f^*(X)} \Big| D_n\Big] \\
&\leq 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big] \mathbb{E}\Big[(\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X) \neq f^*(X)})^2 \Big| D_n\Big]} && \text{(Cauchy-Schwarz ineq.)} \\
&= 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big] \mathbb{P}\Big(\hat{f}_{\hat{\eta}}(D_n; X) \neq f^*(X) \Big| D_n\Big)}. && (3)
\end{aligned}$$

The Cauchy-Schwarz inequality namely reads for any random variables X and Y: $|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$, so $\mathbb{E}(XY) \leq \sqrt{|\mathbb{E}(XY)|^2} \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$. The final equality follows from the fact that the expectation of an indicator function is equal to the probability of the condition of the indicator function being satisfied (which in turn follows from the definition of an expectation).

(f) Because P is a zero-error distribution, we have $f^*(X) = Y$ almost surely and $\inf_{f \in \mathcal{F}} \mathcal{R}_P(f) := \mathcal{R}_P^* = 0$ (see (a)). Thus:

$$\begin{aligned}
\mathbb{P}\Big(\hat{f}_{\hat{\eta}}(D_n; X) \neq f^*(X) \Big| D_n\Big) &= \mathbb{P}\Big(\hat{f}_{\hat{\eta}}(D_n; X) \neq Y \Big| D_n\Big) \\
&= \mathbb{E}\left[\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X) \neq Y} \Big| D_n\right] \\
&= \mathcal{R}_P(\hat{f}_{\hat{\eta}}(D_n)) \\
&= \mathcal{R}_P(\hat{f}_{\hat{\eta}}(D_n)) - \mathcal{R}_P^* \\
&:= \ell(f^*, \hat{f}_{\hat{\eta}}(D_n)).
\end{aligned}$$

(g) Plugging in the result from the previous question, so $\mathbb{P}\Big(\hat{f}_{\hat{\eta}}(D_n;X) \neq f^*(X)\Big|D_n\Big) = \ell(f^*,\hat{f}_{\hat{\eta}}(D_n))$, into the inequality (3), we obtain:

$$\begin{aligned} & \ell(f^*,\hat{f}_{\hat{\eta}}(D_n)) \le 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\Big|D_n\Big]\ell(f^*,\hat{f}_{\hat{\eta}}(D_n))} \\ \iff \quad & \sqrt{\ell(f^*,\hat{f}_{\hat{\eta}}(D_n))} \le 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\Big|D_n\Big]} \\ \iff \quad & \ell(f^*,\hat{f}_{\hat{\eta}}(D_n)) \le 4\mathbb{E}\Big[\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\Big|D_n\Big]. \end{aligned}$$

(h) The bound in the lecture was:

$$\ell(f^*,\hat{f}_{\hat{\eta}}(D_n)) \le 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\Big|D_n\Big]}.$$

Hence the upper bound of the lecture is the square root of the upper bound found above. Because we are working with 0-1 cost here, we know that any meaningful upper bound of the excess risk is smaller than one. In other words, the bound we found above suggests a lower excess risk for the plug-in classifier than the bound found in the lecture (which was not based on the zero-error assumption).

3. (a) For $h = 1/2$ this corresponds to the zero-error assumption in question 2, because then $\eta(X) = 0$ or 1 with probability 1. For $h = 0$ no extra restrictions are imposed by the assumption, since $|\eta(X) - 1/2| \ge 0$ is true for any $\eta(X) \in [0,1]$. So the margin condition is more general than the zero-error assumption, because the zero-error assumption is a special case of the margin condition.

(b) This holds because $\mathbb{1}_{|\eta(X)-1/2|<h}$ equals 1 with probability zero due to the margin condition:

$$\mathbb{P}(|\eta(X)-1/2| < h) = 1 - \mathbb{P}(|\eta(X)-1/2| \ge h) = 1 - 1 = 0\,.$$

Therefore:

$$\begin{aligned} \mathbb{E}\Big[|\eta(X)-1/2|\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)}\mathbb{1}_{|\eta(X)-1/2|<h}\Big|D_n\Big] &\le \mathbb{E}\Big[|\eta(X)-1/2|\mathbb{1}_{|\eta(X)-1/2|<h}\Big|D_n\Big] \\ &\le h\ \mathbb{E}\Big[\mathbb{1}_{|\eta(X)-1/2|<h}\Big] \\ &= h\ \mathbb{P}(|\eta(X)-1/2| < h) = 0\,, \end{aligned}$$

where the first inequality uses that the indicator function is smaller or equal than 1, and the second inequality uses that $|\eta(X) - 1/2|\mathbb{1}_{|\eta(X)-1/2|<h} \le h\ \mathbb{1}_{|\eta(X)-1/2|<h}$. The expectation is therefore equal to zero, because it is clearly non-negative (as $[|\eta(X) - 1/2|\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)}\mathbb{1}_{|\eta(X)-1/2|<h} \ge 0$ for any X and D_n).

(c) Starting with the expression for the excess risk:

$$
\begin{aligned}
\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) &= \mathbb{E}\Big[|2\eta(X)-1|\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)}\Big|D_n\Big] \\
&= 2\mathbb{E}\Big[|\eta(X)-1/2|\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)}\Big(\mathbb{1}_{|\eta(X)-1/2|<h}+\mathbb{1}_{|\eta(X)-1/2|\geq h}\Big)\Big|D_n\Big] \\
&= 2\mathbb{E}\Big[|\eta(X)-1/2|\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)}\mathbb{1}_{|\eta(X)-1/2|<h}\Big|D_n\Big] \\
&\qquad + 2\mathbb{E}\Big[|\eta(X)-1/2|\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)}\mathbb{1}_{\big|\eta(X)-1/2\big|\geq h}\Big|D_n\Big]. \\
&= 2\mathbb{E}\Big[|\eta(X)-1/2|\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)}\mathbb{1}_{\big|\eta(X)-1/2\big|\geq h}\Big|D_n\Big]. \qquad (4)
\end{aligned}
$$

Where the final equality follows from (b). Recall that if $\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)$, then $\hat{\eta}(D_n;X)\leq 1/2<\eta(X)$ or $\eta(X)\leq 1/2<\hat{\eta}(D_n;X)$. In other words, then $|\eta(X)-1/2|\leq |\hat{\eta}(D_n;X)-\eta(X)|$ and also $\mathbb{1}_{|\eta(X)-1/2|\geq h}\leq \mathbb{1}_{|\hat{\eta}(D_n;X)-\eta(X)|\geq h}$. So

$$
\begin{aligned}
|\eta(X)-1/2|\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)}\mathbb{1}_{\big|\eta(X)-1/2\big|\geq h} &\leq |\hat{\eta}(D_n;X)-\eta(X)|\mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X)\neq f^*(X)}\mathbb{1}_{\big|\hat{\eta}(D_n;X)-\eta(X)\big|\geq h} \\
&\leq |\hat{\eta}(D_n;X)-\eta(X)|\mathbb{1}_{\big|\hat{\eta}(D_n;X)-\eta(X)\big|\geq h}
\end{aligned}
$$

Combining with (4) this gives the final conclusion:

$$
\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 2\mathbb{E}\Big[|\hat{\eta}(D_n;X)-\eta(X)|\mathbb{1}_{\big|\hat{\eta}(D_n;X)-\eta(X)\big|\geq h}\Big|D_n\Big]
$$

(d) Starting from the inequality in (c), and using the Cauchy-Schwarz inequality gives

$$
\begin{aligned}
\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) &\leq 2\mathbb{E}\Big[|\hat{\eta}(D_n;X)-\eta(X)|\mathbb{1}_{\big|\hat{\eta}(D_n;X)-\eta(X)\big|\geq h}\Big|D_n\Big] \\
&\leq 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\Big|D_n\Big]\mathbb{E}\Big[\big(\mathbb{1}_{\big|\hat{\eta}(D_n;X)-\eta(X)\big|\geq h})^2\Big|D_n\Big]} \qquad \text{(Cauchy-Schwarz ineq.)} \\
&= 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\Big|D_n\Big]\mathbb{P}\Big(|\hat{\eta}(D_n;X)-\eta(X)|\geq h\Big|D_n\Big)}. \qquad (5)
\end{aligned}
$$

(e) Using Markov's inequality:

$$
\mathbb{P}\Big(|\hat{\eta}(D_n;X)-\eta(X)|\geq h\Big|D_n\Big) = \mathbb{P}\Big(\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\geq h^2\Big|D_n\Big) \leq \frac{\mathbb{E}\Big[\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\Big|D_n\Big]}{h^2}
$$

Injecting the last inequality into (5) yields the conclusion:

$$
\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq \frac{2}{h}\mathbb{E}\Big[\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\Big|D_n\Big]
$$

(f) Recall the bounds on the excess risk of the plug-in classifier we found so far:

(i) Under the zero-error condition, $h=1/2$:

$$
\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 4\,\mathbb{E}\Big[\big(\hat{\eta}(D_n;X)-\eta(X)\big)^2\Big|D_n\Big].
$$

(ii) Under the margin condition, $h \in (0, 1/2)$:

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq \frac{2}{h}\, \mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big].$$

(iii) No assumptions on P (lecture), $h = 0$:

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 2\, \sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big]}.$$

The bound found under the margin condition is weaker (i.e. higher) for smaller values of h, which is intuitive, because it the margin condition is less restrictive for smaller values of h. For the case $h \to 1/2$, the bound of the margin condition coincides with the bound we found under the zero-error assumption, which is expected, because for $h = 1/2$ the margin condition coincides with the zero-error condition.

However, as $h \to 0$, the bound we found for the margin condition will go to infinity, for sample size n fixed, even though we would hope to match the bound we found in the lecture (which didn't require any assumptions). In other words, the bound found in this exercise is not always an improvement to the bound found in the lectures that was not based on the margin condition. So there is a mis-match between the bound found under the margin condition for $h \to 0$ and the one we found in the lecture. Apparently, if h is too small, so if a very unrestrictive version of the margin condition is used, the bound found in this exercise may not be useful. [The reason is that we used Markov's inequality to derive a bound under the margin condition, which is a rather rough upper bound. If h is really small, than you can imagine that the upper bound found by Markov's inequality can be quite uninformative, as it might even be greater than 1.].

(g) (This question is less important) Note that alternatively you could write $\mathbb{E}[(\hat{\eta}(D_n; X) - \eta(X))^2] = O(n^{-1})$ and work with that, if you are more comfortable with that notation.

Start with (i): Taking expectation on both sides of the inequality of part (e) and invoking the law of iterated expectations:

$$\mathbb{E}\Big[\ell(f^*, \hat{f}_{\hat{\eta}}(D_n))\Big] \leq \frac{2}{h}\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2\Big] \underset{n \to +\infty}{\sim} \frac{2c}{hn},$$

so the plug-in learning rule is weakly consistent. On the other hand, **not assuming the margin condition**, we obtained in the lecture that

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big]}.$$

Taking expectation on both sides, and using the fact that $\mathbb{E}[\sqrt{|Z|}] \leq \sqrt{\mathbb{E}[|Z|]}$ because of the concavity of the square-root function (reverse Jensen's inequality),

$$\begin{aligned}
\mathbb{E}\Big[\ell(f^*, \hat{f}_{\hat{\eta}}(D_n))\Big] &\leq 2\mathbb{E}\left[\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big]}\right] \\
&\leq 2\sqrt{\mathbb{E}\left[\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big]\right]} \\
&= 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2\Big]} \\
&\underset{n \to +\infty}{\sim} 2\sqrt{\frac{c}{n}},
\end{aligned}$$

where the equality follows from the law of iterated expectations.

Under the margin condition, i.e. if $\mathbb{P}(Y = 1|X)$ is bounded away from $1/2$ for all X, the average excess risk of $\hat{f}_{\hat{\eta}}$ tends to zero at least as fast as $O(n^{-1})$. On the other hand, if the margin condition does not hold, i.e. if $\mathbb{P}(Y = 1|X)$ can be arbitrarily close to $1/2$, we can only say that the average excess risk tends to zero at least as fast as $O(n^{-1/2})$, which is much slower. To guarantee a given performance level, we will need many more examples if the margin condition does not hold.

Note however: these are just upper bounds, so $\mathbb{E}\left[\ell(f^*, \hat{f}_{\hat{\eta}}(D_n))\right]$ can in both cases also go to zero at a strictly faster rate.