Midterm

Kevin Chang

October 28, 2025

1

Suppose we'd like to build a classification rule for a finite population Ω of pairs (x_i, y_i) where x_i are integers and y_i are assumed to be in $\{-1, 1\}$. Let n denote the size of Ω.

(i) Compute the function $\hat{y}_{me}(x)$ that minimizes the number of classification errors on this population.

The function that minimizes the number of classification errors (majority vote) is

$$\hat{y}_{\text{me}}(x) = \text{sign}\Big(\sum_{i\in\Omega} \mathbb{1}(x_i = x)\, y_i\Big)$$

(with arbitrary tie-breaking when the sum is zero). The minimal empirical error rate is

$$\frac{1}{n}\sum_{x\in\Omega} \min(n_x^+, n_x^-),$$

where $n_x^+ = |\{x_i \in \Omega \wedge x_i = x : y_i = 1\}|$ and $n_x^- = |\{x_i \in \Omega \wedge x_i = x : y_i = -1\}|$.

(ii) Suppose we instead choose to minimize the square-loss over all possible functions f:

$$R_{sq}[f] = \frac{1}{2n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$$

Minimizing with respect to $f(x)$ separately for each x gives

$$f_{\text{sq}}(x) = \begin{cases} \dfrac{1}{n_x}\displaystyle\sum_{i\in S_x} y_i & n_x > 0, \\ 0, & n_x = 0, \end{cases}$$

where $S_x = \{x_i \in \Omega \wedge x_i = x\,\}, n_x = |S_x|$

(iii) Define the classification rule

$$\hat{y}_{sq}(x) = \begin{cases} 1 & f_{sq}(x) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Compute the fraction of times that $\hat{y}_{sq}$ makes an error. How does it compare to the error rate of $\hat{y}_{sq}$?

Define

$$\hat{y}_{\text{sq}}(x) = \begin{cases} 1, & f_{\text{sq}}(x) > 0, \\ -1, & f_{\text{sq}}(x) < 0, \\ \text{either } \pm 1, & f_{\text{sq}}(x) = 0. \end{cases}$$

Since $\text{sign}(f_{\text{sq}}(x)) = \text{sign}(\sum_{i\in S_x} y_i)$, we have

$$\hat{y}_{\text{sq}}(x) = \hat{y}_{\text{me}}(x).$$

Hence, both classifiers achieve the same empirical error rate:

$$\boxed{\frac{1}{n}\sum_{x} \min(n_x^+, n_x^-)}.$$

2

Let $S = \{(x_i, y_i)\}$ be a set of example-label pairs with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Suppose we want to minimize the empirical loss

$$\frac{1}{2n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2$$

with respect to $w \in \mathbb{R}^d$. Assume the x_i span $\mathbb{R}^d$ and there is a solution w_* that achieves zero loss. Analyze the performance of stochastic gradient descent at minimizing this loss. Assume that w is initialized equal to 0. At each iteration, a stochastic gradient is computed from a single example from S, sampled independently with replacement. Let (x_{i_k}, y_{i_k}) denote the example-label pair selected at iteration k. Assume the step size is a constant.

(i) Show that the SGD iterates take the form

$$w_{k+1} = w_* + A_k(w_k - w_*).$$

For the squared loss, the stochastic gradient is

$$\nabla \ell_k(w_k) = (\langle w_k, x_{i_k} \rangle - y_{i_k}) x_{i_k}.$$

Since $\langle w_*, x_i \rangle = y_i$, we have

$$\nabla \ell_k(w_k) = \langle w_k - w_*, x_{i_k} \rangle x_{i_k}.$$

Thus,

$$w_{k+1} = w_k - \eta \nabla \ell_k(w_k) = w_k - \eta \langle w_k - w_*, x_{i_k} \rangle x_{i_k},$$

or equivalently,

$$\boxed{w_{k+1} = w_* + A_k(w_k - w_*), \quad A_k = I - \eta x_{i_k} x_{i_k}^\top.}$$

(ii) Compute $\mathbb{E}[A_k]$.

Since i_k is drawn uniformly,

$$\boxed{\mathbb{E}[A_k] = I - \eta G, \quad G := \tfrac{1}{n} \sum_{i=1}^{n} x_{i_k} x_{i_k}^\top.}$$

Because $\{x_{i_k}\}$ span $\mathbb{R}^d$, G is symmetric positive definite with eigenvalues $0 < \lambda_{\min} \leq \lambda_{\max}$.

(iii) Show that there exists a step size such that

$$||E[w_T - w_*]|| \leq \beta^T ||w_*||$$

for some $\beta < 1$.

Taking expectations,

$$\mathbb{E}[w_{k+1} - w_*] = (I - \eta G)\, \mathbb{E}[w_k - w_*] \quad \Rightarrow \quad \mathbb{E}[w_T - w_*] = (I - \eta G)^T (w_0 - w_*) = -(I - \eta G)^T w_*.$$

Hence,

$$\|\mathbb{E}[w_T - w_*]\| \leq \|I - \eta G\|_2^T \, \|w_*\| = \big(\max_j |1 - \eta \lambda_j|\big)^T \|w_*\|.$$

Choosing $0 < \eta < 2/\lambda_{\max}$ ensures $\beta := \|I - \eta G\|_2 < 1$, yielding

$$\boxed{\|\mathbb{E}[w_T - w_*]\| \leq \beta^T \|w_*\|, \quad \beta < 1.}$$

3

You want to build a simple classifier on a d-dimensional space, which uses the rule

$$f(x) = \begin{cases} 1 & x_i \geq \alpha \text{ and } x_j \geq \beta \\ -1 & \text{otherwise} \end{cases}$$

That is, the classifier examines exactly two dimensions and declares the example to be positive if the values in those dimensions are both large enough. You are given three sets that are guaranteed to be i.i.d. samples from a larger set S:

(i) S_{train} with n_{train} data points.

(ii) S_{test} with n_{test} data points.

(iii) S_{val} with n_{val} data points.

You use the set S_{train} to find the best parameters α and β for each possible choice of dimensions i and j. This yields a set of predictors $\hat{f}_{ij}$. You select the $\hat{f}_{ij}$ that has the lowest error on S_{test}. Call the selected predictor $\hat{f}_{best}$.

Let E_{test} denote the 0-1 classification error of $\hat{f}_{best}$ on S_{test} and E_{val} denote the 0-1 classification error of $\hat{f}_{best}$ on S_{val}. Provide a bound on the probability that $|E_{test} - E_{val}|$ will be less than ϵ.

Since both 0-1 classification errors are empirical means of i.i.d. Bernoulli trials bounded in $[0, 1]$, by Hoeffding's inequality,

$$\Pr\big[\, |E_{\text{test}} - \mathbb{E}[E_{\text{test}}]| \geq t \,\big] \;\leq\; 2\exp(-2n_{\text{test}}t^2).$$

Setting $t = \sqrt{\frac{\log(2/\epsilon)}{2n_{\text{test}}}}$ gives

$$\Pr\big[\, |E_{\text{test}} - \mathbb{E}[E_{\text{test}}]| \geq t \,\big] \leq \epsilon.$$

Hence, with probability at least $1 - \epsilon$,

$$\boxed{|E_{\text{test}} - \mathbb{E}[E_{\text{test}}]| \leq \sqrt{\frac{\log(2/\epsilon)}{2n_{\text{test}}}}.}$$

Analogously, since E_{val} is an independent empirical estimate of the same underlying risk,

$$\boxed{|E_{\text{test}} - E_{\text{val}}| \;\leq\; 2\sqrt{\frac{\log(2/\epsilon)}{2n_{\text{min}}}}, \quad n_{\text{min}} = \min(n_{\text{test}}, n_{\text{val}}),}$$

with probability at least $1 - \epsilon$.

4

Consider a classification problem where you have access to K models that were trained by hyperparameter tuning on a train-test split (S_{train}, S_{test}). Here, the data sets are assumed to be representative examples for a classification problem. Suppose someone gives you a set S_{val} that they assert was generated to be identically distributed to S_{test}. Let $E_{test}[f_k]$ and $E_{val}[f_k]$ denote the error of f_k on S_{test} and S_{val} respectively.

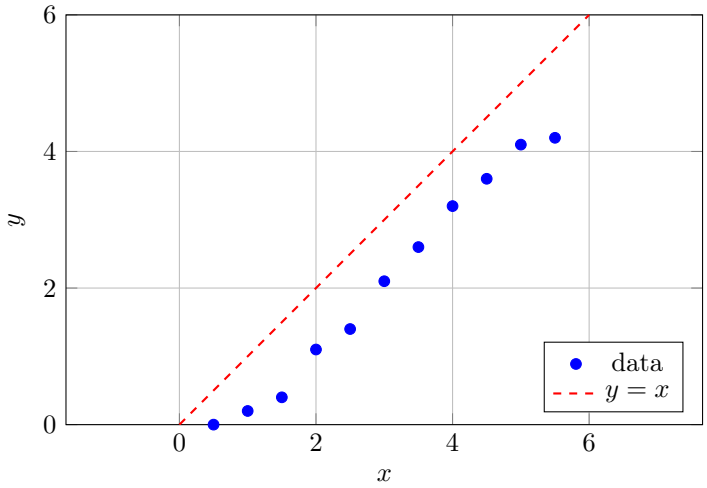
(i) Make a hypothetical scatter plot of points that represents what you'd expect the relationship between E_{test} (x-axis) and E_{val} (y-axis) to look like. Explain your reasoning for why your plot looks the way it does.

The points $(E_{test}[f_k], E_{val}[f_k])$ should lie close to the diagonal line $y = x$, indicating that the models perform similarly on S_{test} and S_{val} since both are assumed to be identically distributed.

However, if S_{val} is *not independent* of S_{test} (e.g., shares correlated samples or overlapping data sources), this dependence can lead to degraded or biased performance estimates. In such cases, the apparent correlation between E_{test} and E_{val} may weaken, and one set may appear artificially optimistic or pessimistic relative to the other.

Hence, we expect:

- A tight, positively correlated cluster of points near the diagonal $y = x$.
- Slight downward deviation ($E_{val} < E_{test}$).



we assume that S_{val} is identically distributed to S_{test}; however not independently.

(ii) Describe conditions under which you'd expect $E_{test}[f_k]$ and $E_{val}[f_k]$ to be nearly equal.

The quantities $E_{test}[f_k]$ and $E_{val}[f_k]$ will be nearly equal when the following conditions hold:

1. S_{val} is drawn *independently and identically distributed* (i.i.d.) from the same distribution as S_{test}.
2. The model selection or hyperparameter tuning does *not* depend on S_{test}.
3. The evaluation protocol (metrics, preprocessing, randomness) is consistent across both sets.