HW1

Kevin Chang

October 14, 2025

Terminology

- System state Y: an unknown random variable.
- Measurement X: an observed random variable statistically related to Y.
- Estimator $\hat{Y}(X)$: a random variable defined as a function of X.
- Probability:
  - Prior: $P[Y]$
  - Posterior: $P[Y \mid X]$
  - Likelihood: $P[X \mid Y]$
- Objective (Risk):

$$R[\hat{Y}] = \mathbb{E}\big[loss(\hat{Y}(X), Y)\big]$$

- Optimal Estimator (Posterior form):

$$\hat{Y}(x) = \mathbb{1}\left\{ P[Y = 1 \mid X = x] \;\geq\; \frac{loss(1,0) - loss(0,0)}{loss(0,1) - loss(1,1)}\, P[Y = 0 \mid X = x] \right\}$$

  - Proof:

$$\mathbb{E}[loss(\hat{Y}(X), Y)] = \int_{-\infty}^{\infty} \mathbb{E}[loss(\hat{Y}(X), Y) \mid X = x] f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} \Big(\mathbb{E}[loss(\hat{Y}(X), 1) \mid X = x]\; P[Y = 1 \mid X = x] + \mathbb{E}[loss(\hat{Y}(X), 0) \mid X = x]\; P[Y = 0 \mid X = x] \Big) f_X(x)\, dx$$

  - Thus, $\hat{Y}(x)$ is chosen according to the label (0 or 1) that minimizes the conditional expected loss.

- Optimal Estimator (Likelihood ratio form):

$$\hat{Y}(x) = \mathbb{1}\left\{ \frac{p(x \mid Y = 1)}{p(x \mid Y = 0)} \;\geq\; \frac{p_0 \left(loss(1,0) - loss(0,0) \right)}{p_1 \left(loss(0,1) - loss(1,1) \right)} \right\}$$

  - Proof by rearrangement of the posterior condition.
  - This corresponds to a *likelihood ratio test*.

Types of errors and successes

- True Positive Rate: $P[\hat{Y} = 1 | Y = 1]$
- False Negative Rate: $P[\hat{Y} = 0 | Y = 1]$
- False Positive Rate: $P[\hat{Y} = 1 | Y = 0]$
- True Negative Rate: $P[\hat{Y} = 0 | Y = 0]$
- Precision: $P[Y = 1 | \hat{Y} = 1]$

Receiver Operating Characteristic(ROC) curve

- Example

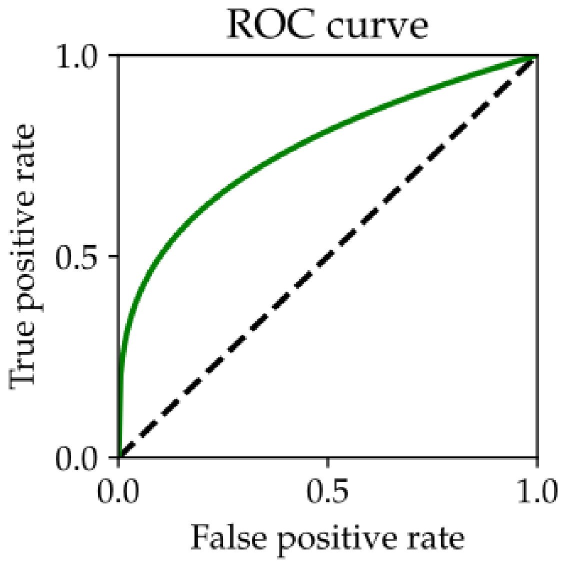

Figure 1: The ROC curve is plotted in the FPR–TPR plane.

- Lemma 2 (Neyman–Pearson Lemma) Suppose the likelihood functions $p(x \mid y)$ are continuous. Then the optimal probabilistic predictor that maximizes TPR subject to an upper bound on FPR is a deterministic likelihood ratio test.
- Properties
  - always passes through $(0, 0)$ and $(1, 1)$,
  - must lie above the main diagonal,
  - is concave.

Fairness

- Key statistical measures include:
  - **Acceptance rate:** $\Pr[\hat{Y} = 1]$
  - **Error rates:** $\Pr[\hat{Y} = 0 \mid Y = 1]$, $\Pr[\hat{Y} = 1 \mid Y = 0]$
  - **Conditional outcome frequency:** $\Pr[Y = 1 \mid R = r]$
- Standard fairness criteria are:
  - **Independence:** $R \perp A$ (equal acceptance rates across groups)
  - **Separation:** $R \perp A \mid Y$ (equal error rates across groups)
  - **Sufficiency:** $Y \perp A \mid R$ (equal outcome frequencies given R)
- It is well known that any two criteria are mutually exclusive in general, except in degenerate cases; thus enforcing one typically precludes the others.

1 Supervised Learning

Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ denote a labeled dataset with $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. For a predictor $f : \mathcal{X} \to \mathcal{Y}$, the *empirical risk* is

$$R_S[f] = \frac{1}{n} \sum_{i=1}^{n} loss\big(f(x_i), y_i\big),$$

Three fundamental questions arise:

- **Representation:** Which function class $\mathcal{F}$ should we select?
- **Optimization:** How can the corresponding learning problem be solved efficiently?
- **Generalization:** How well does the predictor extend from training data to unseen samples?

Perceptron Algorithm The perceptron iteratively updates a weight vector $w \in \mathbb{R}^d$:

- Initialize $w^{(0)} = 0$.
- For $t = 0, 1, 2, \ldots$:
  - Select $i \in \{1, \ldots, n\}$ uniformly at random.
  - If $y_i \langle w^{(t)}, x_i \rangle < 1$, set
  $$w^{(t+1)} = w^{(t)} + y_i x_i,$$
  else $w^{(t+1)} = w^{(t)}$.

Connection to Empirical Risk Minimization The perceptron update can be viewed as stochastic gradient descent (SGD) on Hinge loss:

$$\min_w \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{hinge}}(y_i, \langle w, x_i \rangle) + \|w\|_2^2.$$

- **Hinge loss:**
$$\ell_{\text{hinge}}(y, \hat{y}) = \max\{1 - y\hat{y}, 0\},$$
- **Squared loss:**
$$\ell_{\text{sq}}(y, \hat{y}) = \tfrac{1}{2}(y - \hat{y})^2,$$
- **Logistic loss:**
$$\ell_{\log}(y, \hat{y}) = \begin{cases} -\log(\sigma(\hat{y})), & y = 1, \\ -\log(1 - \sigma(\hat{y})), & y = -1, \end{cases}$$
where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid.

Margin Analysis

- For $w \in \mathbb{R}^d$, define the *margin* on dataset S as
$$\gamma(S, w) = \min_{1 \le i \le n} \frac{|\langle x_i, w \rangle|}{\|w\|}, \qquad \gamma(S) = \max_w \gamma(S, w).$$
- Let $D(S) = \max_{1 \le i \le n} \|x_i\|$.
- Theorem: If S is linearly separable, the perceptron algorithm makes at most $\frac{\left(2+D(S)^2\right)}{\gamma(S)^2}$ margin mistakes.
- *Proof sketch.* Expanding the update yields
$$\|w^{(t+1)}\|^2 = \|w^{(t)} + y_i x_i\|^2 = \|w^{(t)}\|^2 + 2y_i \langle w^{(t)}, x_i \rangle + \|x_i\|^2 \le \|w^{(t)}\|^2 + 2 + D(S)^2.$$
Meanwhile, progress in the margin direction ensures
$$\langle w^*, w^{(t+1)} - w^{(t)} \rangle \ge \gamma(S),$$
for an optimal separator w^*, leading to the stated bound.

Generalization Bound Let S_n be n i.i.d. samples from a distribution $\mathcal{D}$ admitting a perfect linear separator. Let $w(S_n)$ denote the perceptron's output after convergence on S_n, and let $(X, Y) \sim \mathcal{D}$ be independent of S_n. Then

$$P\big[Yw(S_n)^T X < 1\big] \leq \mathbb{E}[\frac{2 + D(S_{n+1})^2}{(n+1)\,\gamma(S_{n+1})^2}],$$

where $D(S_{n+1})$ and $\gamma(S_{n+1})$ are defined analogously on $S_{n+1} = S_n \cup \{(X, Y)\}$.

2 Representation

- **Lifting functions $\Phi(x)$:** Transform a given set of features into a more expressive feature space.
- **Common strategies:**
  - **Template matching:** For example, $x_0 = \max\{v^\top x, 0\}$, which can be interpreted as a sliding window that activates when a feature satisfies certain conditions.
  - **Polynomial features:** In d dimensions with maximum degree p, the number of monomial coefficients is $\binom{d+p}{p}$.
- **Dimensionality:** How high must the lifted dimension be?

  To gain intuition, stack n data points $x_1, \dots, x_n \in \mathbb{R}^d$ into a matrix $X \in \mathbb{R}^{n \times d}$, where each row corresponds to a sample. Predictions over the dataset can then be expressed as

  $$\hat{y} = Xw.$$

  If the x_i are linearly independent and $d \geq n$, then any prediction vector y can be realized by an appropriate weight vector w. Thus, feature design often aims to lift data into sufficiently high-dimensional spaces so that the feature matrix X has linearly independent columns, enabling greater expressivity.
- **Kernels**
  - Given a lifting functoin Φ, the kernel function is

    $$k(x, z) := \Phi(x)^\top \Phi(z),$$

    which ensures that for any $x_1, \dots, x_n$, the Gram matrix K with entries $K_{ij} = k(x_i, x_j)$ is positive semidefinite.
  - A function f can be expressed as

    $$f(x) = w^\top \Phi(x) = \sum_{1 \leq i \leq n} \alpha_i \, k(x_i, x).$$

  - Moreover, if k_1 and k_2 are kernels, then both $k_1 k_2$ and $k_1 + k_2$ are valid kernels.

3 Optimization

- **Gradient Descent.**
  - *Procedure.*
    * Minimize the empirical loss

      $$\phi(w) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i, w), y_i).$$

    * Initialize $w_0 \in \mathbb{R}^d$.

* For $t = 0, 1, 2, \dots$:

$$w_{t+1} = w_t - \alpha_t \frac{1}{n} \sum_{i=1}^{n} \nabla \mathcal{L}(f(x_i, w), y_i), \quad \alpha_t > 0.$$

- *Theorem.*
  * A vector v is a descent direction for ϕ at w_0 if

$$\phi(w_0 + tv) < \phi(w_0) \quad \text{for some } t > 0.$$

  * A point w^* is a local minimizer only if $\nabla\phi(w^*) = 0$.
  * If $\phi : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex, then

$$w^* \text{ is a global minimizer of } \phi \iff \nabla\phi(w^*) = 0.$$

- **Stochastic Gradient Descent (SGD).**
  - *Procedure.*
    * Minimize the empirical loss

$$\phi(w) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i, w), y_i).$$

    * Initialize $w_0 \in \mathbb{R}^d$.
    * For $t = 0, 1, 2, \dots$, sample $i \in \{1, \dots, n\}$ uniformly at random and update

$$w_{t+1} = w_t - \alpha_t \nabla_w \mathcal{L}(f(x_i, w_t), y_i), \quad \alpha_t > 0.$$

  - *Remark.* SGD reduces to the perceptron algorithm when applied with the hinge-type loss

$$\ell(y, \hat{y}) = \max(-y\hat{y}, 0),$$

    using a linear predictor $f(x, w) = w^\top x$.
  - *Analysis.*
    * Assume SGD update rule is given by

$$w_{t+1} = w_t - \alpha_t g_t(w_t; \eta_t),$$

      ,where $g_t(w_t, \eta_t) = \nabla_w \mathcal{L}(f(x_t, w_t), y_t)$ is a stochastic gradient computed from a sample $\eta_t = (x_t, y_t)$.
    * Assume the gradient is bounded:

$$\|g_t(w_t; \eta_t)\| \le B, \quad \forall t.$$

    * We expand the squared norm of the distance to the optimum w_*:

$$\|w_{t+1} - w_*\|^2 = \|w_t - w_*\|^2 - 2\alpha_t \langle g_t(w_t; \eta_t), w_t - w_* \rangle + \alpha_t^2 \|g_t(w_t; \eta_t)\|^2.$$

    * Taking expectations and using the law of iterated expectation gives

$$\mathbb{E}[\langle g_t(w_t; \eta_t), w_t - w_* \rangle] = \mathbb{E}[\langle \nabla \mathcal{L}(w_t), w_t - w_* \rangle].$$

    * Summing from $t = 0$ to $T - 1$ and rearranging terms yields

$$\sum_{t=0}^{T-1} \alpha_t \mathbb{E}[\langle \nabla \mathcal{L}(w_t), w_t - w_* \rangle] \le \frac{1}{2} \|w_0 - w_*\|^2 + \frac{B^2}{2} \sum_{t=0}^{T-1} \alpha_t^2.$$

    * By convexity of $\mathcal{L}$,

$$\mathcal{L}(w_t) - \mathcal{L}(w_*) \le \langle \nabla \mathcal{L}(w_t), w_t - w_* \rangle.$$

* Hence,

$$\sum_{t=0}^{T-1} \alpha_t \mathbb{E}[\mathcal{L}(w_t) - \mathcal{L}(w_*)] \leq \frac{\|w_0 - w_*\|^2}{2} + \frac{B^2}{2} \sum_{t=0}^{T-1} \alpha_t^2.$$

* Defining the weighted average iterate

$$\tilde{w}_T = \frac{\sum_{t=0}^{T-1} \alpha_t w_t}{\sum_{t=0}^{T-1} \alpha_t},$$

* and applying convexity again, we obtain the standard SGD convergence bound:

$$\mathbb{E}[\mathcal{L}(\tilde{w}_T) - \mathcal{L}(w_*)] \leq \frac{\|w_0 - w_*\|^2 + B^2 \sum_{t=0}^{T-1} \alpha_t^2}{2 \sum_{t=0}^{T-1} \alpha_t}.$$

4 Generalization

- The goal is to bound the difference between the *empirical risk* $R_S[f]$ (measured on a sample) and the *true risk* $R[f]$ (expected loss under the underlying distribution).
- **Hoeffding's Inequality.** For independent random variables $Z_1, \ldots, Z_n$ bounded in $[a_i, b_i]$,

$$P[\bar{Z} - \mathbb{E}[\bar{Z}] \geq t] \leq \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \quad \bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i.$$

If the loss $\mathcal{L}$ is bounded in $[0, 1]$, then for any f,

$$P[R_S[f] > R[f] + t] \leq e^{-2nt^2}.$$

- **Finite Hypothesis Class.** Applying the union bound to a finite hypothesis set $\mathcal{F}$ yields, with probability at least $1 - \delta$,

$$|R_S[f] - R[f]| \leq \sqrt{\frac{\ln |\mathcal{F}| + \ln(1/\delta)}{2n}}, \quad \forall f \in \mathcal{F},$$

where $\ln |\mathcal{F}|$ measures the *complexity* of the model family. The generalization gap thus scales as $\mathcal{O}\left(\sqrt{\frac{\text{complexity}(\mathcal{F})}{n}}\right)$.