

# HW1

Kevin Chang

September 26, 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:  $P[Y | X]$
  - Likelihood:  $P[X | Y]$

- Objective (Risk):

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

- Optimal Estimator (Posterior form):

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

- Proof:

$$\begin{aligned} \mathbb{E}[\text{loss}(\hat{Y}(X), Y)] &= \int_{-\infty}^{\infty} \mathbb{E}[\text{loss}(\hat{Y}(X), Y) | X = x] f_X(x) dx \\ &= \int_{-\infty}^{\infty} (\mathbb{E}[\text{loss}(\hat{Y}(X), 1) | X = x] P[Y = 1 | X = x] + \mathbb{E}[\text{loss}(\hat{Y}(X), 0) | X = x] P[Y = 0 | X = x]) f_X(x) dx \end{aligned}$$

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

- Optimal Estimator (Likelihood ratio form):

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

- 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

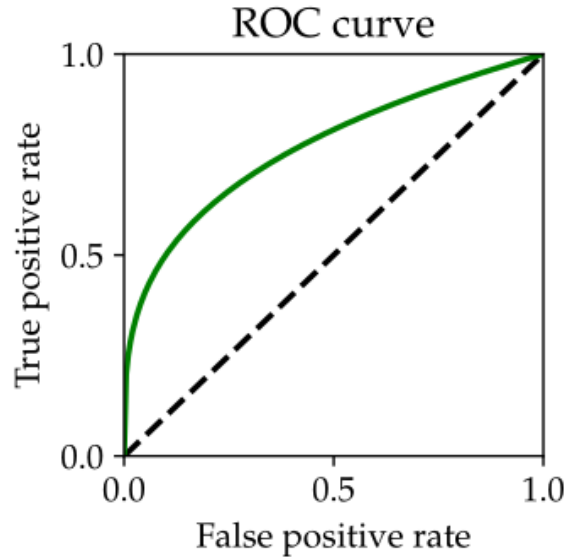


Figure 1: The ROC curve is plotted in the FPR–TPR plane.

- Lemma 2 (Neyman–Pearson Lemma) Suppose the likelihood functions  $p(x | y)$  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:**  $\Pr[\hat{Y} = 0 | Y = 1]$ ,  $\Pr[\hat{Y} = 1 | Y = 0]$
  - **Conditional outcome frequency:**  $\Pr[Y = 1 | R = r]$
- Standard fairness criteria are:
  - **Independence:**  $R \perp A$  (equal acceptance rates across groups)
  - **Separation:**  $R \perp A | Y$  (equal error rates across groups)
  - **Sufficiency:**  $Y \perp A | R$  (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  $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$  denote a labeled dataset with  $x_i \in \mathcal{X}$  and  $y_i \in \mathcal{Y}$ . For a predictor  $f : \mathcal{X} \rightarrow \mathcal{Y}$ , the *empirical risk* is

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

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  $t = 0, 1, 2, \dots$ :
  - Select  $i \in \{1, \dots, n\}$  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:**

$$\ell_{\text{sq}}(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2,$$

- **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

$$\gamma(S, w) = \min_{1 \leq i \leq n} \frac{|\langle x_i, w \rangle|}{\|w\|}, \quad \gamma(S) = \max_w \gamma(S, w).$$

- Let  $D(S) = \max_{1 \leq i \leq n} \|x_i\|$ .
- Theorem: If  $S$  is linearly separable, the perceptron algorithm makes at most  $\frac{(2+D(S)^2)}{\gamma(S)^2}$  margin mistakes.
- *Proof sketch.* Expanding the update yields

$$\|w^{(t+1)}\|^2 = \|w^{(t)} + y_i x_i\|^2 = \|w^{(t)}\|^2 + 2y_i \langle w^{(t)}, x_i \rangle + \|x_i\|^2 \leq \|w^{(t)}\|^2 + 2 + D(S)^2.$$

Meanwhile, progress in the margin direction ensures

$$\langle w^*, w^{(t+1)} - w^{(t)} \rangle \geq \gamma(S),$$

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

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

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:** 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.