Lecture 02: Predictive Learning

Qiang Sun

January 2026

Roadmap

1. Supervised learning: prediction + explanation
2. Regression: loss, risk, ERM, overfitting, regularization
3. OLS: solution and model-based interpretation
4. Classification: Bayes rule, decision boundary, logistic regression
5. Other losses: logistic, hinge (SVM), 0-1
6. (Optional extension) feature engineering, neural nets, high-dimensional + ridge

Supervised Learning

Goal

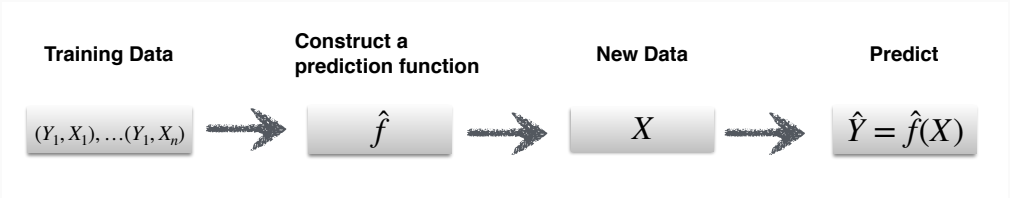
Predict an unknown response Y from features X using labeled data.

- **Regression:** $Y \in \mathbb{R}$ (continuous)
- **Classification:** $Y \in \{1, 2, \ldots, K\}$ (categorical)

Key Idea (Two tasks)

Supervised learning usually combines **prediction** and (often) **explainability/variable selection**.

Generic Workflow (Supervised Learning)



Caption

A generic workflow for supervised learning (data $\rightarrow$ model training $\rightarrow$ evaluation $\rightarrow$ deployment).

Example: Image Classification (Cats vs Dogs)

Setup

We observe labeled images (X_i, Y_i) where $Y_i \in \{\text{cat}, \text{dog}\}$.

We train a classifier $\widehat{f}$ so that for a new image X, we predict $\widehat{Y} = \widehat{f}(X)$.

Key Idea (Generalization)

The key question is how well $\widehat{f}$ predicts on **new** data, not just on training data.

Two Goals of Supervised Learning

1. **Prediction:** Given a new X, predict Y via $\widehat{Y} = \widehat{f}(X)$.
2. **Explainability (variable selection):** Find a small subset of features $X_1, \ldots, X_d$ most related to Y.

Comment

Prediction is the main goal of ML; inference/explanation is central in statistics.

Regression Analysis

Regression Setup: Population vs Sample

Population level

$$(Y, X) \sim P_{Y,X}, \qquad Y \in \mathbb{R},\ X \in \mathbb{R}^d.$$

Sample level

We observe i.i.d. samples

$$(Y_1, X_1), \ldots, (Y_n, X_n) \overset{\text{i.i.d.}}{\sim} P_{Y,X}.$$

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbf{X} = [X_{ij}] \in \mathbb{R}^{n \times d}$.

Notation

(Y_i, X_i) is random; (y_i, x_i) is the observed realization.

Inference vs Prediction (Key Contrast)

Example: $X \sim \mathcal{N}(\theta, 1)$ with unknown θ

We observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$.

- **Inference/statistics:** infer population quantities (e.g., θ)
  - point estimate $\widehat{\theta}_n$, confidence intervals, hypothesis testing
- **Prediction/learning:** predict a new draw X_{n+1} based on $X_{1:n}$
  - focus: generalization performance

Intuition

Regression summarizes the relationship between a response Y and predictors X.

Given data $(Y_i, X_i)_{i=1}^n$, we want a prediction function f so that $f(X)$ is "close" to Y.

Key Idea (Main question)

If $f(X)$ and Y are random, how do we measure their closeness?

Loss Functions and Risk

Let $\mathcal{X}$ be the feature space and $\mathcal{Y}$ be the label space.

Loss function

A loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ measures the cost of predicting $\widehat{Y}$ when the truth is Y.

Expected risk (population risk)

$$R(f) \stackrel{\Delta}{=} \mathbb{E}_{(Y,X)\sim P}\big[\ell(f(X), Y)\big].$$

Examples of Loss Functions (Regression)

Common choices

1. ℓ_2 (OLS): $\ell(f(X), Y) = |f(X) - Y|^2, \quad R(f) = \mathbb{E}|f(X) - Y|^2.$
2. ℓ_1 (LAD): $\ell(f(X), Y) = |f(X) - Y|, \quad R(f) = \mathbb{E}|f(X) - Y|.$

Why ℓ_2 is popular

Mathematically simple, computationally convenient, locally quadratic, and links to Gaussian MLE.

A Fundamental Theorem for ℓ_2 Loss

Theorem

Let

$$f^* = \arg\min_f \mathbb{E}\big[(Y - f(X))^2\big].$$

Then

$$f^*(x) = \mathbb{E}[Y \mid X = x].$$

Key Idea (Regression function)

The conditional mean $g(x) \triangleq \mathbb{E}[Y \mid X = x]$ is the target of ℓ_2 regression.

Proof Sketch (Orthogonality Trick)

Let $\bar{f}(X) = \mathbb{E}[Y \mid X]$. Then:

$$\mathbb{E}(Y - f(X))^2 = \mathbb{E}(Y - \bar{f}(X))^2 + \mathbb{E}(\bar{f}(X) - f(X))^2$$

because the cross term vanishes:

$$\mathbb{E}[(Y - \bar{f}(X))(\bar{f}(X) - f(X))] = 0.$$

Interpretation

The first term is irreducible noise; the second is approximation error minimized at $f = \bar{f}$.

From Population Risk to Empirical Risk

Population risk uses the unknown $P_{Y,X}$, so we minimize the empirical risk:

$$\widehat{R}(f) \triangleq \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2, \qquad \widehat{f} = \arg\min_{f} \widehat{R}(f).$$

A trivial (bad) minimizer

$$\widehat{f}(x) = \begin{cases} Y_i, & x = X_i \\ \text{anything}, & x \neq X_1, \ldots, X_n. \end{cases}$$

Overfitting

A model is too flexible and begins to fit noise instead of signal.

Key Idea (Regularization)

Control model complexity by restricting the hypothesis class or adding penalties/constraints.

Examples of hypothesis classes $\mathcal{F}$

- Linear: $\mathcal{F} = \{f(x) = \beta^\top x\}$
- Polynomial: $\mathcal{F} = \{f(x) = \text{poly}(x)\}$
- Smooth nonparametric: $\mathcal{F} = \left\{f : \int (f'(x))^2 dx < \infty\right\}$
- Neural nets: $\mathcal{F} = \{f : \ f \text{ is a residual network}\}$

Ordinary Least Squares (OLS)

OLS regression

Given $(Y_i, X_i)_{i=1}^n$ with design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$,

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2.$$

Closed form (when $\mathbf{X}^\top \mathbf{X}$ invertible)

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Model-Based Interpretation of OLS

Consider the generative model

$$Y = \beta^\top X + \epsilon, \qquad \mathbb{E}[\epsilon \mid X] = 0, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Key result

Under Gaussian noise, the MLE of β equals the OLS solution:

$$\widehat{\beta}^{\text{MLE}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2.$$

Key Idea (Why model-based?)

It tells you when an estimator is optimal, enables confidence intervals/p-values, and yields a generative story for the data.

Classification

Classification

Categorical response $Y \in \{C_1, \ldots, C_K\}$ (often $Y \in \{1, \ldots, K\}$).

- **Binary classification:** $K = 2$, often $Y \in \{-1, +1\}$.
- **Multiclass:** $K > 2$ (can reduce via one-vs-one / one-vs-rest).

Goal: learn a mapping $h : \mathcal{X} \to \{-1, +1\}$ so that $h(X)$ matches Y.

0–1 Loss and Expected Risk

For $Y \in \{-1, +1\}$, the 0–1 loss is

$$\ell(Y, \widehat{Y}) = \mathbb{I}(Y \neq \widehat{Y}).$$

Expected risk

$$R(h) = \mathbb{E}[\ell(Y, h(X))] = \mathbb{P}(Y \neq h(X)).$$

(Your notes also express ℓ_2 loss as $|Y - h(X)|^2 = 4\mathbb{I}(Y \neq h(X))$.)

Bayes rule

$$h^* = \arg\min_h R(h) = \arg\min_h \mathbb{P}(Y \neq h(X)).$$

Theorem (Bayes classifier)

Let $r(x) \triangleq \mathbb{P}(Y = +1 \mid X = x)$. Then

$$h^*(x) = \begin{cases} +1, & r(x) > \frac{1}{2}, \\ -1, & \text{otherwise.} \end{cases}$$

Key Idea (Core task in classification)

Model/estimate the conditional probability $r(x) = \mathbb{P}(Y = +1 \mid X = x)$.

Decision Boundary

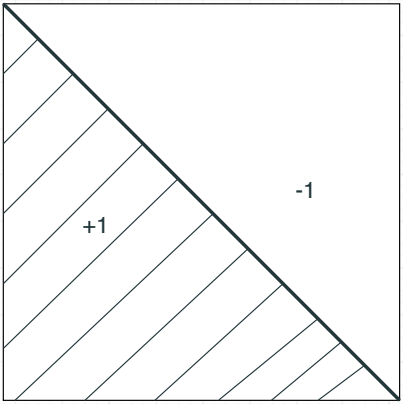
Decision boundary

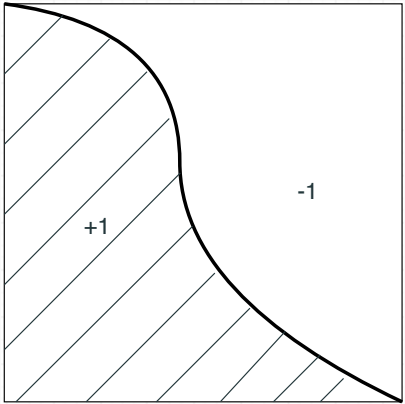
Given $r(x) = \mathbb{P}(Y = +1 \mid X = x)$, define

$$D(r) = \{x : \ r(x) = \tfrac{1}{2}\}.$$

Decision Boundary

- Points with $r(x) > \frac{1}{2}$ fall in the positive region.
- Points with $r(x) < \frac{1}{2}$ fall in the negative region.
- Linear vs nonlinear boundary depends on the form of $r(x)$ (or its score function).





Logistic model

Model the conditional probability by a score function $f(x)$:

$$\mathbb{P}(Y = +1 \mid X = x) = \frac{1}{1 + e^{-f(x)}}, \qquad \mathbb{P}(Y = -1 \mid X = x) = \frac{1}{1 + e^{f(x)}}.$$

Equivalently,

$$\mathbb{P}(Y = y \mid X = x) = \frac{1}{1 + e^{-yf(x)}}, \quad y \in \{-1, +1\}.$$

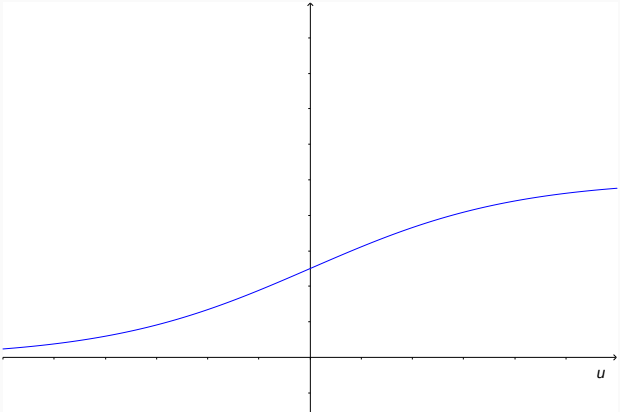
Margin

The quantity $yf(x)$ is the *margin* (we want it large).

Probit model

$$\mathbb{P}(Y = +1 \mid X = x) = \Phi(f(x)),$$

where Φ is the CDF of $\mathcal{N}(0, 1)$.



Logistic Regression as MLE

We treat $p_X(x)$ as a nuisance parameter and focus on $p_f(y \mid x)$:

$$p_f(y \mid x) = \frac{1}{1 + e^{-yf(x)}}.$$

Log-likelihood (up to constants in f)

$$\ell_n(f) = \sum_{i=1}^{n} \log p_f(Y_i \mid X_i) = -\sum_{i=1}^{n} \log\bigl(1 + e^{-Y_i f(X_i)}\bigr).$$

Key Idea (MLE objective)

$$\widehat{f} = \arg\max_f \ell_n(f) \quad \Longleftrightarrow \quad \widehat{f} = \arg\min_f \sum_{i=1}^{n} \log\bigl(1 + e^{-Y_i f(X_i)}\bigr).$$

Overfitting in Logistic Regression

If we do not constrain f, a trivial (bad) solution can separate the training data:

$$\widehat{f}(x) = \begin{cases} +\infty, & X_i = x,\ Y_i = +1, \\ -\infty, & X_i = x,\ Y_i = -1, \\ \text{arbitrary}, & \text{elsewhere.} \end{cases}$$

Key Idea (Regularize the function class)

Prevent overfitting by restricting $f \in \mathcal{F}$ (linear, smooth, RKHS, neural net, etc.).

Examples

- Linear logistic regression: $f(x) = \beta_0 + \beta^\top x$.
- Nonparametric logistic regression: f continuous with $\int [f''(x)]^2 < \infty$.

Risk Minimization View: Logistic Loss

Logistic loss

$$\ell^{\text{logistic}}(y, f(x)) = \log\left(1 + e^{-yf(x)}\right).$$

Logistic risk

$$R(f) = \mathbb{E}\left[\log(1 + e^{-Yf(X)})\right].$$

Log-odds interpretation

Under the logistic model,

$$f(x) = \log \frac{\mathbb{P}(Y = +1 \mid X = x)}{\mathbb{P}(Y = -1 \mid X = x)}.$$

Other Loss Functions for Classification

- ℓ_2 surrogate (as in notes):

$$\sum_{i=1}^{n}(1 - Y_i\beta^\top X_i)^2$$

- 0–1 loss:

$$\mathbb{I}\big(Y\, f(X) < 0\big)$$

- Hinge loss + SVM:

$$\min_{\beta} \sum_{i=1}^{n}(1 - Y_i\beta^\top X_i)_+ + \lambda\|\beta\|_2^2$$

Summary

Likelihood view motivates logistic loss; ERM view allows many losses (logistic, hinge, squared, 0–1).

Loss Functions

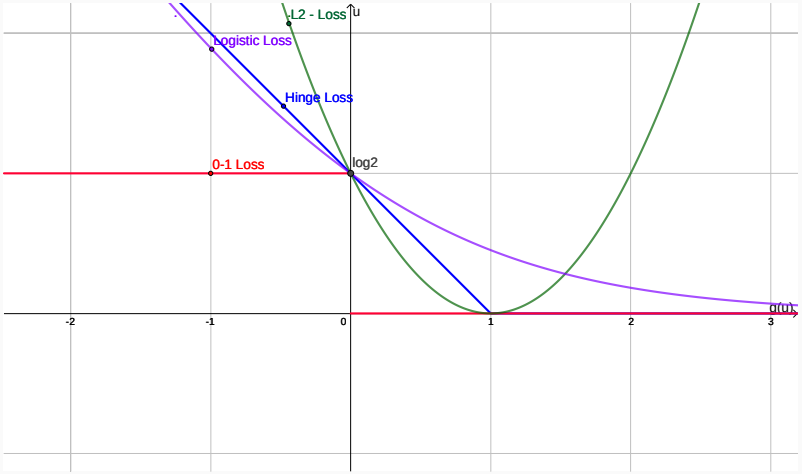


Figure 1: Four losses: logistic, ℓ_2, 0–1, and hinge.

Optional Extensions

Linear Regression Can Be Very Flexible

Linear regression becomes flexible through **feature engineering**:

1. Transformations: $\log(X_1),\ \sqrt{X_1},\ X_1^2$
2. Interactions: $X_j X_k$
3. Basis expansion: $f(X) = \sum_{j=1}^{d} \beta_j h_j(X)$
4. Indicator features: $\mathbb{I}(X_j \in A)$

Categorical Variables and Dummy Coding

Categorical variable

A variable that takes values in a finite set of categories.

Dummy coding

Gender:

$$\mathbb{I}(X = \text{male}) = \begin{cases} 1, & X = \text{male} \\ 0, & X = \text{female} \end{cases}$$

For K categories, use $K - 1$ dummy variables (no ordering assumed).

Neural computation unit / perceptron

$$f(X) = \sigma(\beta^\top X + \beta_0)$$

Common activations: identity, ReLU, sigmoid.

Neural network hypothesis

A family $\mathcal{F}$ of functions obtained by composing such units.

Key Idea (Parametric vs nonparametric)

A neural net with finitely many parameters is a **parametric** model.

Wrap-up

Wrap-Up

- Supervised learning: regression vs classification; prediction vs inference
- Regression: loss $\rightarrow$ risk $\rightarrow$ ERM; overfitting $\rightarrow$ regularization
- OLS: closed form and Gaussian MLE interpretation
- Classification: 0–1 risk; Bayes rule; decision boundary
- Logistic regression: likelihood and ERM (logistic loss); alternatives (hinge/SVM)
- Extensions: feature engineering, neural nets