The Classification Problem and Statistical Framework

Andrew Nobel

September, 2021

Unsupervised vs. Supervised Learning

Unsupervised: Find structure in unlabeled data $x_1, \ldots, x_n$

- PCA and SVD
- Clustering

Supervised: Use labeled data $(x_1, y_1), \ldots, (x_n, y_n)$ to make predictions about an unlabeled sample x

- Classification: response y_i is binary or categorical
- Regression: response y_i is numerical, real-valued

The Classification Problem

Classification

Data: Labeled pairs $(x_1, y_1), \ldots, (x_n, y_n)$ with

- $x_i \in \mathcal{X}$ space of *predictors* (often $\mathcal{X} \subseteq \mathbb{R}^d$)
- $y_i \in \{0, 1\}$ response or *class label*

Goal: Given an *unlabeled* predictor $x \in \mathcal{X}$, assign it to class 0 or 1

- Classification of examples may be of financial or scientific importance
- Obtaining labels may be expensive or difficult

Idea: Use labeled examples to classify unlabeled ones

Example: Spam Recognition

Predictor: $x =$ vector of features extracted from text of email, e.g.,

- presence of keywords ("cheap", "cash", "medicine")
- presence of key phrases ("Dear Sir/Madam")
- use of words in all-caps ("VIAGRA")
- point of origin of email

Response: $y = 1$ if email is spam, $y = 0$ otherwise

Task: Given sample $(x_1, y_1), \ldots, (x_n, y_n)$ of labeled emails, construct a prediction rule to classify future email messages as spam or not-spam

Examples

Medical Testing

- x contains the (numerical) results of d diagnostic tests
- $y = 1$ if patient is at risk for a disease, $y = 0$ if not

Loan Default Prediction

- x contains features related to credit history of loan applicant
- $y = 1$ if applicant defaults, $y = 0$ if applicant repays loan

Overview

- Prediction rules, decision regions, and zero-one loss
- Classification in a stochastic setting
- Optimality: Bayes rule and the Bayes risk

Measuring Errors in Prediction

Definition: A *prediction rule* is a map $\phi : \mathcal{X} \to \{0, 1\}$. Regard $\phi(x)$ as a prediction of the class label associated with x

Zero-One loss: Performance of ϕ on pair (x, y) given by

$$\ell(\phi(x), y) \;=\; \mathbb{I}(\phi(x) \neq y) \;=\; \begin{cases} 1 & \text{if } \phi(x) \neq y \\ 0 & \text{if } \phi(x) = y \end{cases}$$

Summary table. Four possible outcomes: two correct, two errors

	$\phi(x) = 1$	$\phi(x) = 0$
$y = 1$	correct (1,1)	error (1,0)
$y = 0$	error (0,1)	correct (0,0)

Decision Regions and Decision Boundary

Note: Every rule $\phi : \mathcal{X} \to \{0, 1\}$ partitions $\mathcal{X}$ into two sets

$$
\begin{aligned}
\mathcal{X}_0(\phi) &= \{x \in \mathcal{X} : \phi(x) = 0\} \\
\mathcal{X}_1(\phi) &= \{x \in \mathcal{X} : \phi(x) = 1\}
\end{aligned}
$$

Terminology

- Sets $\mathcal{X}_0(\phi), \mathcal{X}_1(\phi)$ called *decision regions* of ϕ
- Interface between $\mathcal{X}_0(\phi)$ and $\mathcal{X}_1(\phi)$ called *decision boundary* of ϕ

Classification Problem Revisited

Picture

- Write sample $(x_1, y_1), \ldots, (x_n, y_n)$ as points $x_i \in \mathcal{X}$ with labels y_i
- Look for decision regions that (mostly) separate zeros and ones

Two Related Issues

- Tradeoff between complexity and separation
- Will selected rule perform well on future, unlabeled, samples?

The Stochastic Setting

Stochastic Setting

Assumptions

- Observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ random
- (X_i, Y_i) drawn independently from distribution P on $\mathcal{X} \times \{0, 1\}$
- Future observation (X, Y) drawn independently from same distribution P

Key Stochastic Quantities

1. Prior probabilities of $Y = 0$ and $Y = 1$
2. Conditional probability of $Y = 1$ given $X = x$
3. Conditional distribution of X given $Y = 0$ and $Y = 1$

Prior Probabilities

Given: Joint pair $(X, Y) \in \mathcal{X} \times \{0, 1\}$

Define: Prior probabilities $\pi_0 = \mathbb{P}(Y = 0)$ and $\pi_1 = \mathbb{P}(Y = 1)$

Notes

- Probability of seeing class $Y = 0$ or $Y = 1$ *prior* to observing X
- π_0, π_1 represent relative abundance of class 1 and 1
- Note that $\pi_0 + \pi_1 = 1$
- Cases in which $\pi_1 >> \pi_0$ or vice versa can be difficult

Unconditional and Conditional Densities of X

Given: Joint pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$

Define: Unconditional and conditional densities of X

- $f(x) =$ *unconditional density* of X

$$\mathbb{P}(X \in A) \;=\; \int_A f(x)\, dx \qquad A \subseteq \mathcal{X}$$

- $f_0(x), f_1(x) =$ *class-conditional densities* of X

$$\mathbb{P}(X \in A \,|\, Y = y) \;=\; \int_A f_y(x)\, dx \qquad A \subseteq \mathcal{X}$$

Note: f_0 and f_1 tell us about separability of 0s and 1s

Conditional Distribution of Y Given X

Given: Joint pair $(X, Y) \in \mathcal{X} \times \{0, 1\}$

Define: Conditional probability $\eta(x) \; = \; \mathbb{P}(Y = 1 \,|\, X = x)$

- Posterior probability that $Y = 1$ given that $X = x$
- Note that $\mathbb{P}(Y = 0 \,|\, X = x) = 1 - \eta(x)$.

Regimes:

- $\eta(x) \approx 1 \;\Rightarrow Y$ is likely to be 1 given $X = x$
- $\eta(x) \approx 0 \;\Rightarrow Y$ is likely to be 0 given $X = x$
- $\eta(x) \approx 1/2 \;\Rightarrow$ value of Y uncertain given $X = x$

Relations Among Distributions

1. By the law of total probability we have

$$f(x) \;=\; \pi_0 f_0(x) \,+\, \pi_1\, f_1(x)$$

Moreover, as f_0 and f_1 are densities $\int f_0(x)dx = \int f_1(x)dx = 1$

2. By Bayes theorem we know

$$\eta(x) \;=\; \frac{\pi_1 f_1(x)}{f(x)} \;=\; \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) \,+\, \pi_1\, f_1(x)}$$

Risk of a Prediction Rule

Recall: Performance of rule ϕ on single pair (x, y) given by zero-one loss

$$\ell(\phi(x), y) \;=\; \mathbb{I}(\phi(x) \neq y) \;=\; \begin{cases} 1 & \text{if } \phi(x) \neq y \\ 0 & \text{if } \phi(x) = y \end{cases}$$

Definition: The *risk* of a fixed prediction rule ϕ on a random pair (X, Y) is its *expected loss*

$$R(\phi) \;=\; \mathbb{E}[\mathbb{I}(\phi(X) \neq Y)] \;=\; \mathbb{P}(\phi(X) \neq Y)$$

which is just the probability that ϕ misclassifies X

Optimality and the Bayes Rule

Bayes Rule and Bayes Risk

Definition: The *Bayes Rule* ϕ^* for the pair (X, Y) is

$$\phi^*(x) = \operatorname*{argmax}_{k=0,1} \mathbb{P}(Y = k \mid X = x)$$

- $\phi^*(x)$ is the most likely value of Y given $X = x$
- $\phi^*(x)$ depends on distribution of (X, Y), usually unknown

Definition: The *Bayes risk* R^* for (X, Y) is the risk of the Bayes rule

$$R^* \ = \ R(\phi^*) \ = \ \mathbb{P}(\phi^*(X) \neq Y)$$

Optimality of the Bayes Rule

Note: For binary Y the Bayes Rule has the form

$$\phi^*(x) \;=\; \begin{cases} 1 & \text{if } \eta(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Theorem: The Bayes rule ϕ^* for (X, Y) is optimal: for every classification rule $\phi : \mathcal{X} \to \{0, 1\}$ we have $R^* \leq R(\phi)$.

Fact: The Bayes risk R^* can be written in the form

$$R^* = \mathbb{E} \min\{\eta(X), 1 - \eta(X)\}$$

Understanding the Bayes Risk

Fact: Let $(X, Y) \in \mathcal{X} \times \{0, 1\}$ be a jointly distributed pair

1. Bayes risk $R^* \in [0, 1/2]$
2. $R^* = 0$ iff $\eta(x) \in \{0, 1\}$ iff Y is a function of X
3. $R^* = 1/2$ iff $\eta(x) \equiv 1/2$ which implies that $Y \perp\!\!\!\perp X$
4. If $Y \perp\!\!\!\perp X$ then $\phi^*(x)$ is constant (1 if $\pi_1 \geq \pi_0$ and 0 if $\pi_0 < \pi_1$)

Fixed vs. Data Dependent Prediction Rules

Observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ iid $\sim (X, Y)$

Fixed rule $\phi : \mathcal{X} \to \{0, 1\}$

- $\phi(x)$ predicts class label of x without regard to D_n
- Risk $R(\phi) = \mathbb{P}(\phi(X) \neq Y)$ is a constant

Data-dependent rule $\hat{\phi} : \mathcal{X} \times (\mathcal{X} \times \{0, 1\})^n \to \{0, 1\}$

- $\hat{\phi}(x) = \hat{\phi}(x : D_n)$ predicts class label of x *based on* D_n
- Risk $R(\hat{\phi}) = \mathbb{P}(\hat{\phi}(X) \neq Y \mid D_n)$ is a random variable