Machine Learning, STOR 565

Classification: Problem and Stochastic Framework

Andrew Nobel

February, 2020

Unsupervised and Supervised Learning

Unsupervised learning: Unlabeled data $x_1, \ldots, x_n$

- SVD and PCA
- Clustering
- Mixture modeling

Supervised learning: Labeled data $(x_1, y_1), \ldots, (x_n, y_n)$ consisting of (predictor, response) pairs

- Classification: response $y_i \in \{-1, 1\}$
- Regression: response $y_i \in \mathbb{R}$

Classification

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

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

Goal: Given an *unlabeled* predictor x, assign it to class -1 or $+1$

Definition: A *prediction rule* is a map $\phi : \mathcal{X} \to \{-1, +1\}$.

- $\phi(x) =$ prediction of the class label associated with x

Classification: Motivation

Problem of assigning unlabeled object to one of two groups arises in many circumstances

- Predictors often readily available, relatively inexpensive/easy to obtain
- Response not readily available, relatively expensive/difficult to obtain

Understanding and modeling the relationship between the predictors and the response is of scientific interest

- Is a simple (linear or quadratic) model sufficient?
- If predictor is high dimensional, are only a few components enough?

Examples

Medical Testing

- $x \in \mathbb{R}^d$ contains the (numerical) results of d diagnostic tests
- $y = 1$ if patient is at risk for a disease, $y = -1$ if not

Object Recognition

- $x \in \mathbb{R}^d$ contains the pixel intensities from a satellite image
- $y = 1$ if image contains a man-made object, $y = -1$ otherwise

Loan Default Prediction

- $x \in \mathbb{R}^d$ contains features data to credit history of loan applicant.
- $y = 1$ if applicant pays back loan in full, $y = -1$ if applicant defaults

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, $y = -1$ otherwise

Key Issues

- How to measure the loss/error of a prediction
- Placing the classification problem in a stochastic setting
- How to assess the overall performance of a prediction rule
- Identifying the optimal rule and its performance
- How to finding good prediction rules from observations

Measuring the Loss of a Prediction

Question: How to assess the performance of a rule $\phi : \mathcal{X} \to \{-1, +1\}$ on an observed pair (x, y)?

Common to use the **Zero-One Loss Function**

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

Note: Two types of errors

$$\phi(x) = 1, y = 0 \;\text{ and }\; \phi(x) = 0, y = 1$$

given equal weight

Measuring the Loss of a Prediction, cont.

Note: The Zero-One Loss can be written equivalently in terms of $\phi(x) \cdot y$ as

$$\ell(\phi(x), y) = \begin{cases} 1 & \text{if } \phi(x) \cdot y < 0 \\ 0 & otherwise \end{cases}$$

In this more general form the decision rule $\phi : \mathcal{X} \to \{-1, +1\}$ can be replaced by a general function $f : \mathcal{X} \to \mathbb{R}$.

Decision Regions and Decision Boundary

Every decision rule $\phi : \mathcal{X} \to \{-1, +1\}$ partitions the predictor space $\mathcal{X}$ into two sets called **decision regions**

$$\begin{aligned} \mathcal{X}_+(\phi) &= \{x \in \mathcal{X} : \phi(x) = +1\} \\ &= \text{points } x \text{ assigned by } \phi \text{ to } +1 \end{aligned}$$

and

$$\begin{aligned} \mathcal{X}_-(\phi) &= \{x \in \mathcal{X} : \phi(x) = -1\} \\ &= \text{points } x \text{ assigned by } \phi \text{ to } -1 \end{aligned}$$

The boundary between the regions $\mathcal{X}_+(\phi)$ and $\mathcal{X}_+(\phi)$ is called the **decision boundary** of ϕ.

Classification Problem Revisited

Picture

- View the sample $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, +1\}$ as a set of labeled points in $\mathcal{X}$ with x_i having label y_i.
- Look for a simple prediction rule (a partition of $\mathcal{X}$ into two sets) that separates the points labeled -1 from those labeled $+1$

Key Issues:

- Tradeoff between the complexity of the partition and its ability to separate the the -1s and $+1$s
- How well is the selected rule likely to perform on future, unlabeled, samples?

Stochastic Setting

Assumptions

- The available data are independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a fixed distribution P on $\mathcal{X} \times \{-1, +1\}$.
- Future observations will be drawn from the same distribution P.

Notation: (X, Y) denotes a generic pair with distribution P, independent of the observations

Key Quantities

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

Prior Probabilities of Y

Define prior probabilities $\pi_1 = \mathbb{P}(Y = +1)$ and $\pi_{-1} = \mathbb{P}(Y = -1)$

- Probability of seeing class $Y = -1$ or $Y = +1$ *prior* to observing x
- π_1, π_{-1} represent relative abundance of class -1 and $+1$
- Note that $\pi_1 + \pi_{-1} = 1$
- Cases in which $\pi_{-1} >> \pi_1$ or vice versa can be difficult

Unconditional and Conditional Densities of X

Assume: $X \in \mathcal{X} \subseteq \mathbb{R}^d$ has unconditional density $f(x)$, that is,

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

Define: For $y \in \{-1, 1\}$ let $f_y(x)$ be the **class-conditional density** of X given $Y = y$

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

Note: f_1 and f_{-1} tell us about the separability of -1s and $+1$s.

Conditional Distribution of Y Given X

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

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

Regimes:

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

Relations Among Distributions

The law of total probability: $f(x) \;=\; \pi_{-1} f_{-1}(x) \,+\, \pi_1\, f_1(x)$

Bayes theorem

$$\eta(x) \;=\; \frac{\pi_1 f_1(x)}{f(x)} \;=\; \frac{\pi_1 f_1(x)}{\pi_{-1} f_{-1}(x) \,+\, \pi_1\, f_1(x)}$$

Expected Loss

Recall: The 0/1 loss of decision rule $\phi : \mathcal{X} \to \{-1, +1\}$ is given by

$$\ell(\phi(x), y) \;=\; \mathbb{I}(\phi(x) \neq y)$$

Measure performance of a decision rule ϕ by its *expected loss* (risk)

$$R(\phi) \;=\; \mathbb{E}[\ell(\phi(X), Y)]$$

Important: Note that

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

is just the probability that ϕ misclassifies a sample.