STAT 535

10/18/22

Lecture 6

(Q1 today)
HW2 t.b. posted

Loss functions
Expected, Bayes' Loss
Bias and Variance

Lecture II: Prediction – Basic concepts

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

October, 2022

Parametric vs non-parametric ✓

Generative and discriminative models for classification ✓

Generative classifiers

Discriminative classifiers

Generative vs discriminative classifiers

Loss functions

Bayes loss

Variance, bias and complexity

Reading HTF Ch.: 2.1–5,2.9, 7.1–4 bias-variance tradeoff, Murphy Ch.: 1., 8.6[1], Bach Ch.:

[1]Neither textbook is close to these notes except in a few places; take them as alternative perspectives or related reading

The "learning" problem

- **Given**
- a problem (e.g. recognize digits from $m \times m$ gray-scale images)
- a **sample** or (**training set**) of **labeled data**

$$\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$$

  drawn i.i.d. from an unknown P_{XY}
- **model class** $\mathcal{F} = \{f\}$ = set of predictors to choose from

- **Wanted**
- a predictor $f \in \mathcal{F}$ that performs well on future samples from the same P_{XY}
  - "choose a predictor $f \in \mathcal{F}$" = training/learning
  - "performs well on future samples" (i.e. f **generalizes** well) – how do we measure this? how can we "guarantee" it?
  - choosing $\mathcal{F}$ is the **model selection problem** – about this later

A zoo of predictors

- Linear regression
- Logistic regression
- Linear Discriminant (LDA)
- Quadratic Discriminant (QDA)
- CART (Decision Trees)
- K-Nearest Neighbors
- Nadaraya-Watson (Kernel regression)
- Naive Bayes
- Neural networks/Deep learning
- Support Vector Machines
- Monotonic Regression

Loss functions

The **loss function** represents the cost of error in a prediction problem. We denote it by L, where

true

$$L(y, \hat{y}) = \text{the cost of predicting } \hat{y} \text{ when the actual outcome is } y$$

Note that sometimes the loss depends on x directly. Then we would write it as $L(y, \hat{y}, x)$.
As usually $\hat{y} = f(x)$ or $\mathrm{sgn} f(x)$, we will typically abuse notation and write $L(y, f(x))$.

my prediction

Least Squares (LS) loss

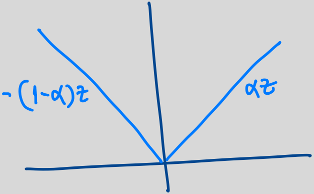
The **Least Squares (LS)** (or **quadratic**) loss function is given by

$$L_{LS}(y, f(x)) = \frac{1}{2}(y - f(x))^2 \tag{5}$$

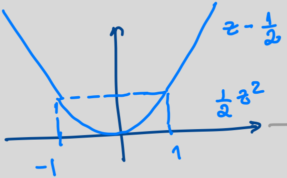
This loss is commonly associated with regression problems.
Example: L_{LS} is the log-likelihood of a regression problem (linear or not) with Gaussian noise.

$$L_{AE}(y, \hat{y}) = |y - \hat{y}|$$

$L_{pinball}$



$$z = y - \hat{y} = y - f(x)$$



$\Rightarrow$ robust estimation

$z = 100$

L_{Huber}

Loss functions for classification

For classification, a natural loss function is the **misclassification error** (also called **0-1 loss**)

$$L_{01}(y, f(x)) = 1_{[y \neq f(x)]} = \begin{cases} 1 & \text{if } y \neq f(x) \\ 0 & \text{if } y = f(x) \end{cases} \quad (6)$$

Sometimes different errors have different costs. For instance, classifying a HIV+ patient as negative (**a false negative** error) incurs a much higher cost than classifying a normal patient as HIV+ (**false positive** error). This is expressed by **asymmetric misclassification costs**. For instance, assume that a false positive has cost one and a false negative has cost 100. We can express this in the matrix

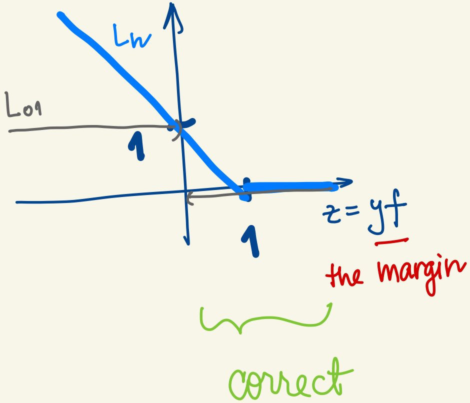
$f(x)$:	+	−
true :+	0	100
−	1	0

In general, when there are p classes, the matrix $L = [L_{kl}]$ defines the loss, with L_{kl} being the cost of misclassifying as l an example whose true class is k.

Special case: $L_{kl} = 1$ iff $k \neq l$

In general: $L_{kk} = 0$

Hinge Loss $\quad L_h(\underset{\pm 1}{y}, \overset{\mathbb{R}}{f}) = \begin{cases} 0 & yf \geq 1 \\ 1 - yf & \text{otherwise} \end{cases}$


Lw
L01
1
1
z = yf
the margin
correct

Expected loss and empirical loss

- Objective of prediction = to minimize expected loss on future data, i.e.

$$\text{minimize } L(f) = E_{P(X,Y)}[L(Y, f(X)] \text{ over } f \in \mathcal{F} \tag{7}$$

We call $L(f)$ above **expected loss**.

Example (Misclassification error $L_{01}(f)$)

$L_{01}(f)$ = probability of making an error on future data.

$$L_{01}(f) = P[Yf(X) < 0] = E_{P_{XY}}[1_{[Yf(X)<0]}] = \Pr[f \text{ makes mistake}] \tag{8}$$

$L_{asymmetric}(f) = ?$

$y \in \pm 1 \qquad L_{+-} \neq L_{-+}$

$L(\mathcal{F}) = \inf_{f \in \mathcal{F}} L(f)$ ← best Loss with $\mathcal{F}$

Ex: $\mathcal{F}_1$ = linear classifiers $\Rightarrow L(\mathcal{F}_1)$

$\mathcal{F}_2$ = quadratic —"— $\Rightarrow L(\mathcal{F}_2)$

$x \in \mathbb{R}^d$

$y \in \{\pm 1\}$

$\mathcal{F}_1 \subset \mathcal{F}_2 \Rightarrow L(\mathcal{F}_2) \leq L(\mathcal{F}_1)$

Empirical loss

$\mathcal{D}_n \sim$ iid P_{xy}

empirical distribution $\hat{P}_{xy}$

$$\hat{L}(f) = E_{\hat{P}_{xy}}[L(y, f)] = \frac{1}{n}\sum_{i=1}^{n} L(y^i, f(x^i))$$ ← can be computed

$$\hat{L}(\mathcal{F}) = \inf_{f \in \mathcal{F}} \hat{L}(f)$$ ← sometimes obtained from training e.g. Linear Regression, L_{LS}

learning algorithms

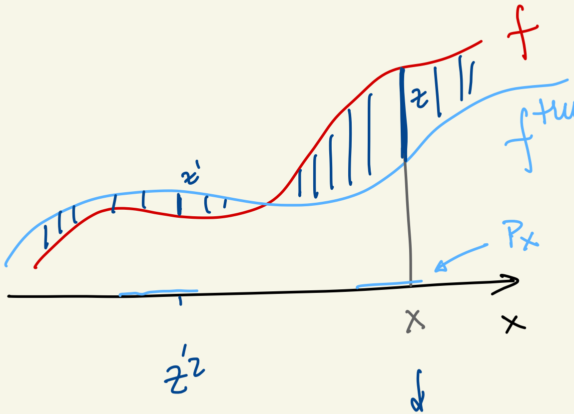
$$\min_{f \in \mathcal{F}} \hat{L}(f)$$ (Linear LS Regression)

exactly or approximately

$$\min_{f \in \mathcal{F}} \underbrace{\hat{L}(f)}_{\text{data}} + \lambda \underbrace{R(f)}_{\text{regularization}}$$ (LASSO, SVM)

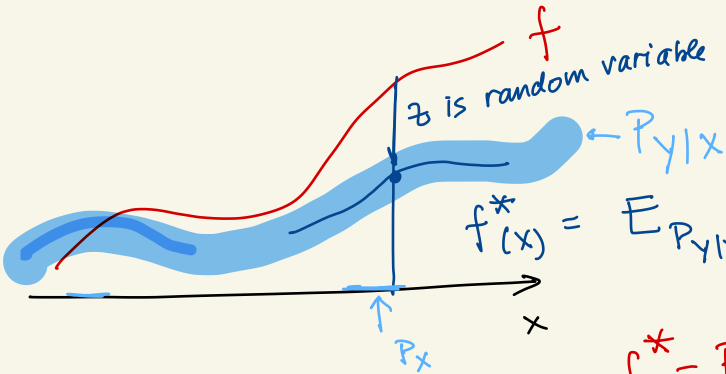
Something else (K-NN, Kernel regression)

$$L_{LS}(f) = E_{P_{xy}}[L_{LS}(y, f(x))]$$



$\Rightarrow$ $y = f^{true}(x)$ deterministic dependence of x

Loss = z^2



$$f^*(x) = E_{P_{y|x}}[Y(x)]$$

best possible predictor for L_{LS}

f^* = Bayes optimal predictor

Expected loss and empirical loss

- Objective of prediction = to minimize expected loss on future data, i.e.

$$\text{minimize}\, L(f) \;=\; E_{P(X,Y)}[L(Y,f(X)] \text{ over } f \in \mathcal{F} \tag{7}$$

  We call $L(f)$ above **expected loss**.
- $L(f)$ cannot be minimized or even computed directly, because we don't know the data distribution P_{XY}.
  Therefore, in training predictors, one uses the **empirical** data distribution given by the sample $\mathcal{D}$.
- The empirical loss (or **empirical error** or **training error**) is the average loss on $\mathcal{D}$

$$\hat{L}(f) \;=\; \frac{1}{n}\sum_{i=1}^{n} 1_{[y^i f(x^i)<0]} \tag{8}$$

- Finally, the value of the **optimal expected loss** for our model class (this is the loss value we are aiming for) is denoted by $L(\mathcal{F})$.

$$L(\mathcal{F}) \;=\; \min_{f\in\mathcal{F}} E_{P(X,Y)}[L(Y,f(X))] \tag{9}$$

  Note that of all the quantities above, we can only know $\hat{L}(f)$ for a finite number of f's in $\mathcal{F}$.

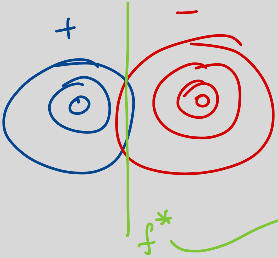
Bayes loss

- How small can the expected loss $L(f)$ be?
  It is clear that
  $$L(\mathcal{F}) \;=\; \min_{f\in\mathcal{F}} L(f) \;\geq\; \min_{f} L(f) \;=\; L^* = L(f^*) \tag{10}$$
  where L^* is taken over all possible functions f that take values in $\mathcal{Y}$.
- L^* is the absolute minimum loss for the given P_{XY} and it is called the **Bayes loss**.
- The Bayes loss is usually not zero

$$f^*(x) = \underset{\hat{y}}{\operatorname{argmin}}\, E_{P_{Y|X}}\, L(y, \hat{y})$$

determ.

true, random



$L^* > 0$

$L(\text{Linear}) = L^*$

LDA

Bayes loss for (binary) classification

- Fix x and assume $P_{Y|X}$ known. Then:
  - Label y will have probability $P_{Y|X}(y|x)$ at this x.
  - No deterministic guess $f(x)$ for y will make the classification error $E_{P_{Y|X=x}}[L_{01}(y, f(x))]$ (unless $P_{Y|X=x}$ is itself deterministic)
  - Best guess minimizes the probability of being wrong. This is achieved by chosing the most probable class
$$y^*(x) \;=\; \operatorname*{argmax}_y P_{Y|X}(y|x) \tag{11}$$
  - The probability of being wrong if we choose $y^*(x)$ is $1 - p^*(x)$, where $p^*(x) = \max_y P_{Y|X}(y|x)$.
- The **Bayes classifier** is $y^*(x)$ as a function of x and its expected loss is the Bayes loss

$$L^*_{01} \;=\; E_{P_X}[1 - p^*(X)] \;=\; E_{P_X}[1 - \max_y P[Y|X]] \tag{12}$$

This shows that the Bayes loss is a property of the problem, via L and P_{XY}, and not of any model class or learning algorithm.

Example

In a classification problem where the class label depends deterministically of the input, the Bayes loss is 0. For example, classifying between written English and written Japanese has (probably) zero Bayes loss.

Example

Consider the least squares loss and the following data distribution: $P_{Y|X} \sim N(g(X), \sigma^2)$. In other words, the Y values are normally distributed around a deterministic function $g(X)$. In this case, optimal least squares predictor is the mean of Y given X, which is equal to $g(X)$. The Bayes loss is the expected squared error around the mean, which is σ^2. Exercise what is the expression of the Bayes loss if $P_{Y|X} \sim N(g(X), \sigma(X)^2)$?

Exercise What is the Bayes loss if (1) $P(Y|X) \sim N((\beta^*)^T X, \sigma^2 I)$ and the loss is L_{LS}; (2) $P(X|Y = \pm 1) \sim N(\mu_\pm, \sigma^2 I)$ and the loss is L_{01} (for simplicity, assume $X \in \mathbb{R}, \mu_{pm} = \pm 1,\ \sigma = 1$); (3) give a formula for the Bayes loss if we know $P(X|Y = \pm 1), P(Y)$, $Y \in \{\pm 1\}$ and the loss is L_{01}. (4) Give an example of a situation when the Bayes loss is 0.