Introduction to Statistical Machine Learning

Xiang Zhou

School of Data Science
Department of Mathematics
City University of Hong Kong

Mathematical Description of Data

Notations

- input – output relation
  - x: inputs, features/attributes, predictors, covariate, factors, independent variables.
  - y: output, response, observatiions, outcomes, dependent variable.
- $\mathcal{X}$ and $\mathcal{Y}$ denote the spaces of the generic x and y variables, respectively.
  - Generally $\mathcal{X} = \mathbb{R}^p$ or $\mathbb{Z}^p$; qualitative features are coded using, for example, dummy variables (such as 0, 1, -1, etc).
  - Typically $\mathcal{Y} \in \mathbb{R}^1$, or takes a finite number of values as a subset of $\mathbb{N}$; it can be a vector in some scenarios.
- (X, Y) denotes the random variable with the joint distribution $p(x, y)$ on the sample space $\mathcal{X} \times \mathcal{Y}$.

Ground truth

- It is usually assumed that the ground truth for the relation between from input to the output is a _____deterministic_____ input-output mapping from $x \in \mathcal{X}$ to $y_{\text{true}} \in \mathcal{Y}$:

$$y_{\text{true}} = f^{\star}(x)$$

  where the ground truth $f^{\star}$ is an unknown function and has to be approximated by learning from the dataset.
- The data/observation is a noisy perturbation of y_{true}.

Data as iid r.v.

- In **supervised learning**, the data (observations, samples) are given as the collection of the pairs

$$\mathtt{D} = \{(x_i, y_i) : 1 \le i \le N\} \subset (\mathcal{X} \times \mathcal{Y})^N$$

which is assumed iid samples of the r.v. [1] (X, Y) with an unknown joint distribution $p(x, y)$ on the product space $\mathcal{X} \times \mathcal{Y}$.

  - Regression: $\mathcal{Y}$ is continuous/numeric, e.g., $\mathbb{R}^1$ or intervals.
  - Classification: $\mathcal{Y}$ is discrete (categorical variable), encoded by integers such as $\{1, \ldots, K\}$. In this case, "y" is usually called "label".

- In **unsupervised learning**, the observations only have $\{x_i\}$, the information y_i is missing or there is no definition of y variable. The task is to identify the pattern of $\{x_i\}$ itself, such as model/dimensionality reduction and clustering.

[1] random variable

Raw data vs features

a remark on "data" defined here and the raw data in data science

- $\mathcal{X}$ may not be the raw data collected from a specific application.
- Raw data is usually quite complex and formally very high dimensional; the direct use of raw data is probably a bad idea in practice.
- The more useful is the feature, only a few (carefully seleced) factors derived from the raw data.
- This process of feature engineering can be done either by <u>domain experts</u> or advanced machine learning methods. [1]

[1] variable selection, model reduction, dimensionality reduction, clustering, interpretable deep learning?, etc.

Measurement error

- The inputs $x^{(i)} \in \mathcal{X}$ are samples from the marginal distribution p_X, i.e., $x^{(i)} \sim X$; in some cases, they are deterministic and assigned by a procedure of experiment design.
- the observed y_i are assumed to be the *perturbed* truth $f^\star(x_i)$ with measurement error ε_i which are assumed to be iid with zero mean and independent from X.
$$y_i = f^\star(x_i) + \varepsilon_i.$$
$\left\{\varepsilon^{(i)}\right\}$ are assumed iid and distributed as a generic r.v. ε.
- This is a convenient model/assumption to specify the joint pdf of (X, Y), even though there might be other types of uncertainty in output observations.
- The effect of the measurement noise ε can never be eliminated by any statistical learning algorithms (irreducible error).

- So, the joint distribution $p_{X,Y}(x, y)$ of (X, Y) is completely determined by the triplet:

$$(p_X, f^\star, p_\varepsilon)$$

  - p_X: the distribution of the input
  - $f^\star$: the input-output function,
  - p_ε: the distribution of measurement error.
- You do not know precisely $f^\star$ and p_ε.
- The joint distribution $p(x, y)$ manifests through the available dataset `D`.

Statistics and machine learning

Different terminologies/jargons:

Machine Learning	**Statistics**
Supervised learning	Classification/regression
Unsupervised learning	Clustering
Semisupervised learning	Classification/regression with missing responses
Features/outcomes	Covariates/responses
Training set/test set	Sample/population
Learner	Statistical model
Generalization error	Misclassification error/prediction error
... ...	... ...

Bayes Rule

Put dataset aside for a while.

Given r.v.s X and Y, find a function $f : \mathcal{X} \to \mathcal{Y}$ so that $f(X)$ can explain Y best in certain sense.

Bayes Rule for regression

conditional expectation as optimal prediction

The best L^2 approximation of a function f of the r.v. X to a r.v. Y is achieved by the conditional probability. The **(generalized) squared error**[1]

$$\mathcal{E}(f) := \mathbb{E}\,|Y - f(X)|^2 \tag{1}$$

has a minimum at

$$\boxed{f^*(x) = \mathbb{E}(Y|X = x)}.$$

i.e.,

$$\mathbb{E}\,|Y - f^*(X)|^2 = \min_{f:\ \text{a Borel function}} \mathbb{E}(|Y - f(X)|^2)$$

$f^*(x) = \mathbb{E}(Y|X = x)$ is called **Bayes rule**.

Note: We did not assume the additive error model here. The applicability of the theorem here is very general.

[1] The jargons "error", "risk", "loss" , even "score" are used exchangeable in many cases

Proof.

(undergraduate prob. course)

- We show first that $\mathbb{E}[(Y - f^*(X))h(X)] = 0$ [a] is true for any function h. Using the double expectation theorem [b], we have

$$\mathbb{E}[(Y - f^*(X))h(X)] = \mathbb{E}\left[\mathbb{E}[Y - f^*(X))|X]h(X)\right]$$
$$= \mathbb{E}\left[\mathbb{E}(Yh(X)|X) - f^*(X)h(X)\right] = 0.$$

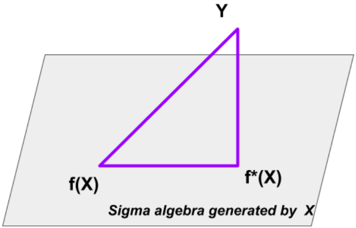
- Note that
$(y - f(x))^2 = (y - f^*(x))^2 + (f(x) - f^*(x))^2 - 2(y - f^*(x))h(x)$
where $h(x) = f(x) - f^*(x)$, then for any f

$$\boxed{\mathbb{E}(|f(X) - Y|^2) = \mathbb{E}(|f^*(X) - Y|^2) + \mathbb{E}\left[|f(X) - f^*(X)|^2\right]} \quad (2)$$

□

[a] sometimes it is denoted $Y - f^*(X) \perp h(X)$, the perpendicular property in L_2 space.

[b] $\mathbb{E}[\mathbb{E}(Y|X)] = \mathbb{E}Y$



reference for elementary math: Understanding Conditional Expectation via Vector Projection

The following exercise is to directly minimize functions in the function space.

Exercise

Use the method of perturbation to solve [a]

$$\inf_f \iint (f(x)-y)^2 p_{X,Y}(x,y)\,dx\,dy$$

where $p_{X,Y}$ *is the joint pdf of the r.v.s* (X,Y)*. The optimal* f^* *satisfies*

$$\int_{\mathcal{Y}} (f^*(x)-y)p_{X,Y}(x,y)\,dy = 0, \qquad \forall x$$

so

$$f^*(x) = \int_{\mathcal{Y}} y \frac{p_{X,Y}(x,y)}{p_X(x)}\,dy = \mathbb{E}[Y|X=x]$$

[a] Rigorously, f is in the p-weighted L_2 space

What if changing the L_2 norm to L_p norm ?

Why L^2 loss

- $\mathbb{E}\, f^*(X) = \mathbb{E}\, Y$: $f^*(X)$ is an <u>unbiased estimate of Y</u>;
- The variance of the difference between Y and the predicted value $f^*(X)$ at $X = x$ is

$$\sigma_*^2(x) := \mathbb{E}\left[(Y - f^*(X))^2 | X = x\right]$$

- Take average of $\sigma^2(x)$ over x, then the averaged uncertainty is

$$\sigma_*^2 := \mathbb{E}_X\, \sigma_*^2(X) = \mathbb{E}\left[|Y - f^*(X)|^2\right] = \mathcal{E}(f^*) = \inf_f \mathcal{E}(f)$$

  This is the variance of the measurement error $Y - f^*(X)$: irreducible error – the error which can not be reduced further.
- For additive measurement error model where $Y = f^\star(X) + \varepsilon$, we have $f^* = f^\star$ and $\sigma_*^2 = \mathrm{Var}(\varepsilon)$.

Bayes Error and Model Error

We have shown in (2) for any two r.v.s X, Y and an arbitrary function f:

$$\begin{aligned}\mathcal{E}(f) &= \underbrace{\mathbb{E}_{X,Y}(|f(X)-Y|^2)}_{\textbf{Mean Square Error}} \\ &= \underbrace{\mathbb{E}_{X,Y}(|f^*(X)-Y|^2)}_{=\mathcal{E}(f^*),\textbf{Bayes error}} + \underbrace{\mathbb{E}_X\left[|f(X)-f^*(X)|^2\right]}_{\textbf{model error}}\end{aligned} \tag{3}$$

where $f^*(x) = \mathbb{E}(Y|X=x)$ is the **Bayes rule**.

- Bayes error: irreducible error;
- Model error: the distance from f to the optimal prediction f^*.

Classification

Next, the same idea, $\inf_f \mathcal{E}(f)$, applied to classification problem...

- Assume $Y \in \{1, \ldots, K\}$ and $X \in \mathbb{R}^p$. So the function $f : \mathcal{X} \to \mathcal{Y}$ we look for is a piece-wise constant $\mathcal{Y}$-valued function; such a function f is better called classifier[1].
- We need a loss function $\ell(Y, f(Y)) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ for penalizing errors due to misclassification.
- Most common choice for classification problem is the **0-1 loss** [2]

$$\ell(Y, f(X)) = I(Y \neq f(X)) := \begin{cases} 1 & \text{if } Y \neq f(X) \\ 0 & \text{if } Y = f(X) \end{cases}.$$

- The expected prediction error(EPR), or generalization error, is then

$$\mathcal{E}(f) = \mathbb{E}\,\ell(Y, f(X)) = \mathbb{P}(Y \neq f(X)) = 1 - \mathbb{P}(Y = f(X)).$$

[1] some references uses the symbol G instead of f

[2] 0-1 loss function here in fact is a K by K identity matrix.

Bayes classifier

$$\begin{aligned}\min_f \mathcal{E}(f) &\Leftrightarrow \max_f \mathbb{P}(Y = f(X)) \\ &= \max_f \int_{\mathcal{X}} \mathbb{P}(Y = f(x)|X = x)p_X(x)\mathrm{d}x \\ &= \int_{\mathcal{X}} \left\{ \max_{f(x)} \mathbb{P}(Y = f(x)|X = x) \right\} p_X(x)\mathrm{d}x\end{aligned}$$

- The Bayes rule minimizing $\mathcal{E}(f)$ is

$$f^*(x) = \operatorname*{argmax}_k \mathbb{P}(Y = k|X = x).$$

Notation for Bayes classifier

- **Bayes classifier**: the maximizer of conditional probability
$$f^*(x) = \operatorname*{argmax}_k \mathbb{P}(Y = k|X = x).$$
- **Bayes error rate**: the minimal value of $\mathcal{E}$
$$\inf_f \mathcal{E}(f) = \mathcal{E}(f^*) = 1 - \mathbb{P}(Y = f^*(X))$$
- **Bayes decision boundary**
The boundary separating the K partition domains in $\mathcal{X}$ on each of which $f^*(x)$ is constant. For the binary classification $(K = 2, \mathcal{Y} = \{-1, 1\})$, the boundary corresponds to the level set where $\mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = -1|X = x) = 0.5$.

Summary of Bayes rule for regression and classification

	regression	classification
$\mathcal{Y} =$	$\mathbb{R}$, continuous	$\{1, \dots, K\}$, categorical
model	$Y = f(X) + \varepsilon$	$\mathbb{P}(Y = k \vert X = x)$
$p_{X,Y}(x, y) =$	$p_X(x)p_\varepsilon(y - f(x))$	$\sum_{k=1}^{K} \pi_k p_X(x; \theta_k)$ (mixture)
loss	mean squared error (L_2)	0-1 loss (misclassification rate)
$f^*(x) =$	$\mathbb{E}(Y \vert X = x)$	$\mathrm{argmax}_k \mathbb{P}(Y = k \vert X = x)$