A.Z
Machine Learning
Nov, 2023

lec 1 Warm-up

Abstract

This part introduces some basic terminology used in Machine Learning, basic ideas about Linear Regression and Binary Classification and some model evaluation methods.

1 Notation

Training sample: $S = \{(x_i, y_i)\}_{i=1}^m$, the training examples $(x, y) \in X \times \mathcal{Y}$ independently drawn from a identical distribution (i.i.d) D defined on $X \times \mathcal{Y}, X$ is a space of inputs, $\mathcal{Y}$ is the space of outputs
Hypothesis h: use to predict outputs given the inputs x. $X \mapsto \mathcal{Y}$
Loss function: L: $L(y, y')$ is the loss incurred when predicting y' when y is true. $\mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}, L(\ldots) \geq 0$
Optimization procedure to find the hypothesis h that minimize the loss on the training sample
Empirical risk: computing the average of the loss on individual instances:

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i)$$

Examples:
- Squared loss is often used in regression:

$$L_{sq}(y, y') = (y' - y)^2, y, y' \in \mathbb{R}$$

- 0/1 loss is often used in classification:

$$L_{0/1}(y, y') = \mathbf{1}_{y \neq y'}$$

- Hamming loss is often used in multilabel learning:

$$L(y, y') = \sum_{j=1}^{d} L_{0/1}(y_j, y'_j), y, y' \in \{-1, +1\}^d$$

True risk, or generalization error:

$$R(h) = \mathbb{E}_{(\mathbf{x}, y) \sim D}[L(h(\mathbf{x}), y)],$$

This is what we would like to minimize.

2 Linear Regression

Training Data: $\{(x_i, y_i)\}_{i=1}^m, (x, y) \in \mathbb{R}^d \times \mathbb{R}$
Loss function: squared loss $L_{sq}(y, y') = (y - y')^2$
Hypothesis class: hyperplanes in $h(\mathbf{x}) = \sum_{j=1}^d w_j x_j + w_0$
Model: $y = h(\mathbf{x}) + \epsilon$, where ϵ is random noise corrupting the output.
We assume zero-mean normal distributed noise: $\epsilon \sim \mathcal{N}(0, \sigma^2)$, with unknown σ
Optimization problem:

$$\text{minimize} \sum_{i=1}^m (y_i - \sum_{j=1}^d w_j x_{ij} + w_0)^2$$
$$\text{w.r.t. } w_j, j = 0, \ldots, d$$

Write this in matrix form:

$$\text{minimize}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$
$$\text{w.r.t. } \mathbf{w} \in \mathbb{R}^{d+1}$$

where:

$$\text{where } \mathbf{X} = \begin{bmatrix} 1 & \mathbf{x}_1 \\ \vdots & \vdots \\ 1 & \mathbf{x}_i \\ \vdots & \vdots \\ 1 & \mathbf{x}_m \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_m \end{bmatrix}$$

Minimum this formula is attained when the derivatives w.r.t $\mathbf{w}$ go to 0:

$$\begin{aligned} & \frac{\partial}{\partial w}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) \\ = & \frac{\partial}{\partial w}\mathbf{y}^T\mathbf{y} - \frac{\partial}{\partial w}2(\mathbf{X}\mathbf{w})^T\mathbf{y} + \frac{\partial}{\partial w}(\mathbf{X}\mathbf{w})^T\mathbf{X}\mathbf{w} \\ = & -2\mathbf{X}^T\mathbf{y} + 2(\mathbf{X}^T\mathbf{X})\mathbf{w} = 0 \end{aligned}$$

If $\mathbf{X}^T\mathbf{X}$ invertible, this formula can be solved by computing a pseudo-inverse matrix.
If not, this formula still can be solved using some other methods.

3 Binary Classification

The label is a binary variable

$$y = \begin{cases} 1 & \text{if } \mathbf{x} \\ 0 & \text{otherwise} \end{cases}$$

For example, we choose as the hypothesis class $\mathcal{H} = \{h : X \mapsto \{0, 1\}\}$ the set of axis prarllel rectangles in $\mathbb{R}^2$, that is

$$h(\mathbf{x}) = (p_1 \leq x_1 \leq p_2) AND (e_1 \leq x_2 \leq e_2)$$

But we don't know the real classification C so we cannot measure exactly how close h is to C
Consistent hypothesis: a hypothesis correctly classifies all training examples
Version space: the set of all consistent hypotheses of the hypothesis class
Most general hypothesis G: cannot be expanded without including negative training examples
Most specific hypothesis S: cannot be made smaller without excluding positive training points

4 Model Evaluation

zero-one loss: $L_{0/1}(y, y') = \mathbf{1}_{y \neq y'}$, where

$$\mathbf{1}_A = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

not a good metric when class distributions are imbalanced or False Negative is costly

4.1 Confusion matrix

True Positives: $m_{TP} =\mid \{\mathbf{x}_i : h(\mathbf{x}_i) = 1 \text{ and } y_i = 1\} \mid$
True Negatives: $m_{TN} =\mid \{\mathbf{x}_i : h(\mathbf{x}_i) = 0 \text{ and } y_i = 0\} \mid$
False Positives: $m_{FP} =\mid \{\mathbf{x}_i : h(\mathbf{x}_i) = 1 \text{ and } y_i = 0\} \mid$
False Negatives: $m_{FN} =\mid \{\mathbf{x}_i : h(\mathbf{x}_i) = 0 \text{ and } y_i = 1\} \mid$
Empirical risk (zero-one loss as the loss function):

$$\hat{R}(h) = \frac{1}{m}(m_{FP} + m_{FN})$$

Precision or Positive Predictive Value:

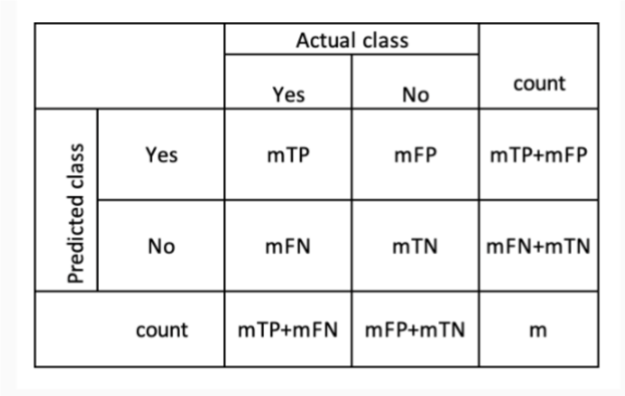
$$\text{Prec}(h) = \frac{m_{TP}}{m_{TP} + m_{FP}}$$

Recall or Sensitivity:

$$\text{Rec}(h) = \frac{m_{TP}}{m_{TP} + m_{FN}}$$

F1 score:

$$F_1(h) = 2\frac{\text{Prec}(h) \cdot \text{Rec}(h)}{\text{Prec}(h) + \text{Rec}(h)} = \frac{2m_{TP}}{2m_{TP} + m_{FP} + m_{FN}}$$



		Actual class		count
		Yes	No	
Predicted class	Yes	mTP	mFP	mTP+mFP
	No	mFN	mTN	mFN+mTN
	count	mTP+mFN	mFP+mTN	m

Figure 1: Confusion Matrix

4.2 Receiver Operating Characteristics (ROC) Curve

x-coordinate: False positive rate $FPR = m_{FP}/m$
y-coordinate: True positive rate $TPR = m_{TP}/m$
The ROC curve is created by plotting TPR and FPR at different thresholds. Each threshold corresponds to a single point on the ROC curve, show the trade-off between TPR and FPR at various threshold levels.
The diagonal line from the bottom left to the top right represents a no-skill classifier (akin to random guessing).

TPR Intuitive Understanding: Imagine a medical test for a specific disease. TPR answers the question: ”Of all the people who actually have the disease, how many did we correctly diagnose as sick?”
TPR Example: If 100 people have a disease, and the test correctly identifies 80 of them as having the disease, the TPR is 80/100 = 0.8 or 80%. This means the test is quite good at catching cases of the disease.
FPR Intuitive Understanding: Using the same medical test, FPR answers the question: ”Of all the people who are actually healthy, how many did we incorrectly diagnose as sick?”
FPR Example : If there are 200 people who don’t have the disease, but the test incorrectly identifies 40 of them as having the disease, the FPR is 40/200 = 0.2 or 20%. This means the test has a tendency to give a fair number of false alarms.

The higher the ROC curve goes, the better the algorithm
If two ROC curves cross it means neither algorithm is globally better
The area under the curve is called AUC

4.3 Evaluation by Testing

We can compute an approximation of the true risk by computing the empirical risk on a independent test sample

$$R_{\text{test}}\,(h) = \sum_{(\mathbf{x}_i, y_i) \in S_{\text{test}}}^{m} L(h(\mathbf{x}_i), y_i)$$

The expectation of $R_{\text{test}}\,(h)$ is the true risk $R(h)$