Classifier

- $D_{Train} : \{\mathcal{X} \times \mathcal{Y}\}^M$
- $D_{Test} : \{\mathcal{X} \times \mathcal{Y}\}^N$
- $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = [C]$ for a C class classification task

A classifier is simply put, a function $h : \mathcal{X} \to \mathcal{Y}$.

Classifier we learn and expect

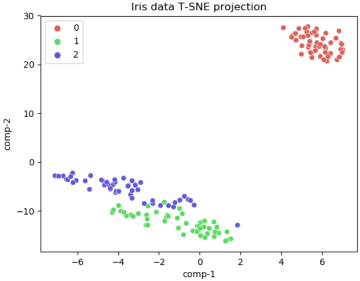
$$\hat{h}(x_i) = y_i \forall (x_i, y_i) \in D_{Train} \tag{1}$$

$$h^*(x_i) = y_i \forall (x_i, y_i) \in D_{Test} \tag{2}$$

$$\tag{3}$$

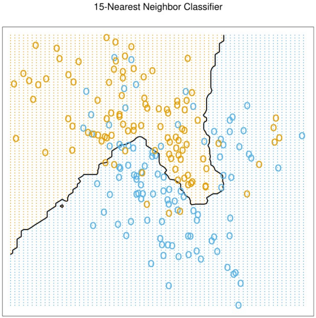
When is $\hat{h} = h^*$?

Most complex $\hat{h}$: Table look-up function



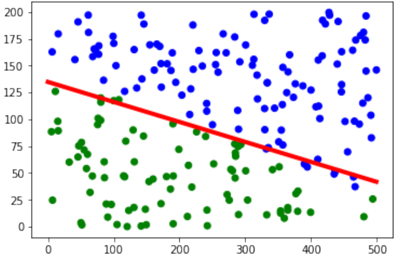
- Can represent any function
- Not usable

Modest $\hat{h}$: Nearest Neighbor Voronoi Tesellation



- This is non-parametric
- Algorithm is not complicated, but inference is!

A Simple $\hat{h}$: Linear Classifier



- Cannot represent all functions
- $\hat{h} = w^T x + b$
- If hj is highly non-linear, gone!

Error

$$\sum_{(x_j, y_j) \in D_{Test}} 1(h(x_j) \neq y_j) \tag{4}$$

Though test error is our target, we cannot learn $\hat{h}$ from test data.

Classification Task

$$\arg\min_{h \in H} \sum_{(x_j, y_j) \in D_{Train}} 1(h(x_j) \neq y_j)$$

Hypothesis Class

For linear functions, $H = \{w \in R^d, b \in R\}$

We always search the best $\hat{h}$ in H

If H is not adequate, then our model cannot generalize on D_{Test}.
i.e. $Error(\hat{h}) >> Error(h^*)$

All constants model

$$c^* = \arg\min_c \sum_{i=1}^{M} 1(c \neq y_i)$$

Linear Hypothesis Class

$$\{w^*, b^*\} = \arg\min_{w,b} \sum_{i=1}^{M} \mathbb{I}(w^T x_i + b \neq y_i)$$

Error function is too stringent

$$\{w^*, b^*\} = \arg\min_{w,b} \sum_{i=1}^{M} |w^T x_i + b - y_i|$$

But our target is discrete [C]

$$\{w^*, b^*\} = \arg\min_{w,b} \sum_{i=1}^{M} \mathbb{I}(\mathrm{sgn}(w^T x_i + b) \neq y_i)$$

Because D_{Train} is scarce, Probabilistic Classifiers often help

$$f(x_i) = \frac{1}{1 + e^{-(w^T x_i + b)}}$$

$$\{w^*, b^*\} = \underset{w,b}{arg\ min} \sum_{i=1}^{M} \mathbb{I}(f(x_i) \neq \frac{y_i + 1}{2})$$

Last proposal

$$\{w^*, b^*\} = \arg\min_{w,b} \sum_{i=1}^{M} \max\left(0, \left(\frac{1}{2} - f(x_i)\right) y_i\right)$$

Qn 1

Assume that we are given a set of features $\{(x_i, y_i) \,|\, i \in \{1, 2, ..., N\}\}$ with $x_i \in R^d$, $y \in \{-1, +1\}$. We wish to train a function $h : R^d \to R$, so that $\text{Sign}(h(x)) = y$. To that aim, we seek to solve the following:

$$\underset{h \in}{\text{minimize}} \sum_{i=1}^{N} [\text{Sign}(h(x_i)) \neq y_i] \tag{5}$$

Moreover, H is the set of all functions that map from R^d to R.

This problem is hard to solve in general. That is why, we resort to several approximations. In the following, mark and explain which ones are good approximator of $I[\text{Sign}(h(x_i)) \neq y_i]$ in Eq. 5.

$$(i) \quad \max\{0, 1 - y_i \cdot h(x_i)\} \quad \text{(Yes/No)} \tag{6}$$

$$(ii) \quad \min\{0, 1 - y_i \cdot h(x_i)\} \quad \text{(Yes/No)} \tag{7}$$

$$(iii) \quad \frac{\exp(-y_i \cdot h(x_i))}{1 + \exp(-y_i \cdot h(_i))} \quad \text{(Yes/No)} \tag{8}$$

$$(iv) \quad \frac{1}{1 + \exp(-y_i \cdot h(x_i))} \quad \text{(Yes/No)} \tag{9}$$

Explanation: ??

Qn 2

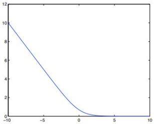
Suppose we restrict $h(x) = w^T x + b$, *i.e.,* $h(x)$ is a linear function. Then write the approximation of the optimization problem defined in Eq. 5 in terms of any (correct) one approximation in the previous question. Specifically, fill up the gaps

$$\text{minimize } \sum_{i=1}^{N} ?? \tag{10}$$

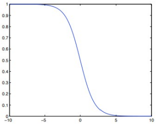
Qn 3

Generally speaking, a classifier can be written as $H(x) = \text{sign}(F(x))$, where $H(x) : \mathbb{R}^d \to \{-1, 1\}$ and $F(x) : \mathbb{R}^d \to \mathbb{R}$. To obtain the parameters in $F(x)$, we need to minimize the loss function averaged over the training set: $\sum_i L(y^i F(x^i))$. Here L is a function of $yF(x)$. For example, for linear classifiers, $F(x) = w_0 + \sum_{j=1}^{d} w_j x_j$, and $yF(x) = y(w_0 + \sum_{j=1}^{d} w_j x_j)$

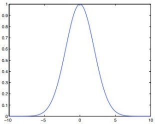
1. [4 points] Which loss functions below are appropriate to use in classification? For the ones that are not appropriate, explain why not. In general, what conditions does L have to satisfy in order to be an appropriate loss function? The x axis is $yF(x)$, and the y axis is $L(yF(x))$.



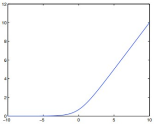
(a)



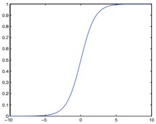
(b)



(c)



(d)



(e)

Qn 4

Consider a Binary classification problem where the dataset D_{Train} is imbalanced. We have 90% examples that belong to class $+1$ and the remaining examples with class -1.

- What is your guess for the best $h \in$ All constants model?
- Compute $Error(h^*) - Error(\hat{h})$ for your guess. Assume that the test set is well-balanced.

Qn 5

Now, let us consider a weighted loss function given by:

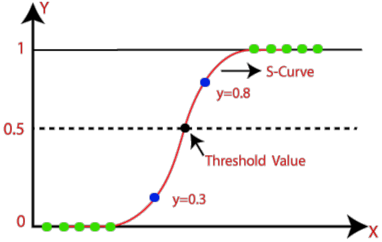
$$\{w^*, b^*\} = \arg\min_{w,b} \sum_{i=1}^{M} r_i \max\left(0, \left(\frac{1}{2} - f(x_i)\right) y_i\right) \tag{11}$$

where $r_i > 0$ are weights associated with loss of each example. Can you propose a weighting scheme for r_i and justify your choice?

Repeat the exercise for the case when test set is also imbalanced with 60% test set examples that belong to class $+1$

Qn 6. Tuning τ in Linear Classification

Recall that Logistic Regression model is given by: $h(x) = \frac{1}{1+e^{-w^T x}}$ where the labels are binary $\mathcal{Y} = \{0, 1\}$



And the loss that we minimize is called *cross-entropy* loss

$$\sum_{(x_j, y_j) \in D_{Train}} -\{y_i \log h(x_i) + (1 - y_i) \log(1 - h(x_i))\} \qquad (12)$$

Finally the decision rule is given by $h(x_i) > 0.5$

contd ...

- Argue that cross entropy loss is a valid loss function.
- What is $||w||$ when training loss is 0. Assume that all features have unit norm $||x|| = 1$
- Is it wrong, if we take $h(x) = \frac{1}{1+e^{+w^T x}}$. Can you tell verbatim, what interpretations change now?

Qn 7. Cheating by using D_{Test}

Now given D_{Test}, the instructor allows you to change the model by modifying the decision rule as $h(x_i) > \tau$ where $\tau \in [0, 1]$. You are free to cheat by inspecting the test set and choosing a τ of your choice. However, you cannot change $\hat{w}, \hat{b}$. Let us evaluate the choices made by the following students:

- Naive student 1: Choose $\tau = 0$
- Naive Student 2: choose $\tau = 1$
- Millennial: choose $\tau = 0.5$
- What would the class choose? Can you pose it as an optimization problem by proposing a loss function and picking τ^* by means of minimizing it?

Qn 8. The meaning of linearity

A function $f(x)$ is said to be linear in x if it satisfies the following two properties

1. $f(x+y) = f(x) + f(y)$
2. $f(\alpha x) = \alpha f(x)$

Are the following equations linear. If yes, then with respect to what parameters?

1. $f(x) = w_1 * x_1 + w_2 * x_2$
2. $f(x) = w_1 * x_1^2 + w_2 * x_2^3$
3. $f(x) = w_1 * \ln x_1 + w_2 * e^{x_2}$
4. $f(x) = x_1 * \ln w_1 + x_2 * e^{w_2}$
5. $f(x) = w^T x \quad w, x \in \mathbb{R}^{\mathrm{d}}$
6. $f(x) = w^T x + b \quad w, x \in \mathbb{R}^{\mathrm{d}} \quad b \in \mathbb{R}$

Qn 9. Minimizing Loss function 1-d case

L-2 Loss in case of linear regression was defined as follows

$$\mathcal{L}_2(w) = \sum_{i=1}^{N} (y_i - wx_i - b)^2$$

$$x_i \in \mathrm{R},\ w \in \mathrm{R},\ b \in \mathrm{R}$$

The interesting thing about linear regression is there exist a closed form solution. This means that the solution can be calculated by minimizing the above function.

Take a gradient of the loss function stated above and prove that the solutions for 1-dimensional case are

$$\widehat{w} = \sum_{i=1}^{N} \frac{(x_i - \overline{x})(y_i - \overline{y})}{(x_i - \overline{x})^2}$$

$$\widehat{b} = \overline{y} - \widehat{w}\overline{x}$$

Qn 10. Regression for general case : Normal equations

L-2 Loss in case of linear regression was defined as follows

$$\mathcal{L}_2(w) = \sum_{i=1}^{N}(y_i - w^T x_i)^2$$

This loss can be neatly written with the help of design matrix X and label vector Y

$$\text{Prove that : } \mathcal{L}_2(w) = ||Xw - Y||^2$$

Now we can take the gradient of the loss function stated above and prove that the solutions for general case. However while taking the gradient a little bit of matrix calculus will be used. We can then finally show that taking the gradient of $\mathcal{L}_2(w)$ and putting it to zero leads us to the normal equations

Derive

$$X^T X w = X^T Y$$

Qn 11. Invertibilty of $X^T X$

Design Matrix $X \in \mathbb{R}^{nXd}$ is a matrix where all samples of the dataset are stacked one below the other. More specifically

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & . & . & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & . & . & x_d^{(2)} \\ x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & . & . & x_d^{(3)} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_1^{(n)} & x_2^{(n)} & x_3^{(n)} & . & . & x_d^{(n)} \end{bmatrix}$$

Here $x_k^{(i)}$ is the k^{th} feature of i^{th} datapoint vector
Recall that the closed form solution of L-2 regression is
$(X^T X)^{-1} X^T Y$
Prove that the inverse of $X^T X$ exist.

Qn 12. Invertibilty of $X^TX + \lambda I$

Although $(X^TX)^{-1}$ does not always exist. $(X^TX + \lambda I)^{-1}$ however does exist. To prove this we will need to understand the definition of positive definite matrices

Given a nxn matrix M The condition for positive definiteness is

$$M \text{ positive-definite} \quad \iff \quad \mathrm{v}^{\mathsf{T}} M \mathrm{v} > 0 \text{ for all } \mathrm{v} \in \mathbb{R}^n \setminus \{0\}$$

A positive definite matrix has a non zero determinant. Therefore its inverse always exists.

Can you prove that $(X^TX + \lambda I)$ is positive definite

Qn 13. MLE for linear Regression

The Linear regression problem can be modelled in a probabilistic way under the assumptions

$$Y_i = w^T x_i + \epsilon_i,$$

$$\epsilon_i \sim \mathrm{N}(0,,\sigma^2)$$

$$Y_i \sim \mathrm{N}(\mathrm{w}^T x_i,,\sigma^2)$$

Prove that the maximising the Likelihood of Data

$$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{n}$$

is equivalent to minimizing the l2-loss that we proposed earlier for the standard regression problem