Lecture 7: Regression(conc.) and Classification

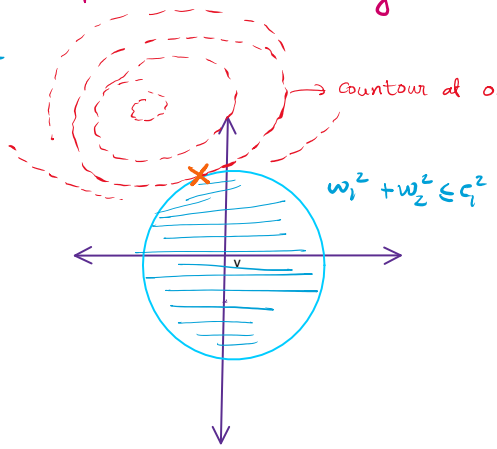
31 January 2024 11:06

Recall the two regression modes:

1. Ridge: $\min \| \underline{\Phi} w - y \|^2 + \lambda \|w\|_2^2$ or, $\min \| \underline{\Phi} w - y \|^2$ sub. $\|w\|_2 \leq c_1$
2. LASSO: $\min \| \underline{\Phi} w - y \|^2 + \lambda \|w\|_1$ or, $\min \| \underline{\Phi} w - y \|^2$ sub $\|w\|_1 \leq c_2$

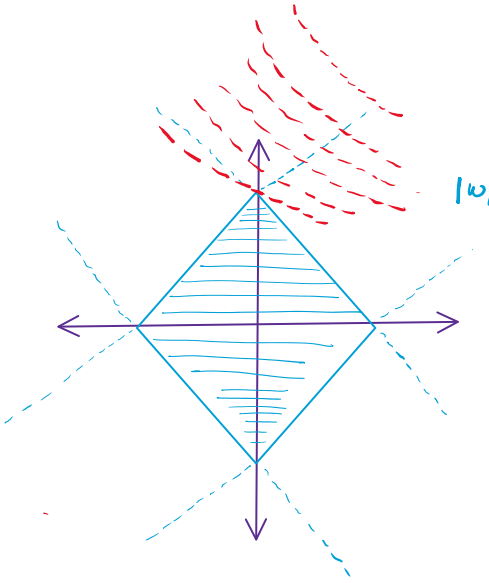
Thus both optimisation problems differ only in constraints.

i) $\|w\|_2 \leq c_2$



If all components are important we use ridge model

ii) $\|w\|_1 \leq c_2$



If only some components are important, we adopt LASSO

REFERENCE: Elements of Statistical Learning by HTF

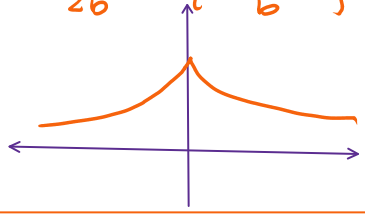
MAP (recall) - prior on w

Gaussian prior ⟶ Ridge

Laplace prior ⟶ LASSO

Laplace $(w_i | \mu, b)$ distribution

$$f(w_i) = \frac{1}{2b} \exp\left\{ \frac{-|w_i - \mu|}{b} \right\}$$



END OF QUIZ-I CONTENT

END OF QUIZ-I CONTENT

CLASSIFICATION:

For regression we'd $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.

For classification we've $y_i \in S$, where S is a finite set.

One simple set S is $\{0,1\}$ (Binary Classification)

Probabilistic Approach: We are interested in probability with which $\hat{x}$ belongs to class $\hat{y}$

we've: $\hat{y} = \underset{y}{\operatorname{argmax}}\, P(Y=y \mid \hat{x})$ $\qquad \hat{x} = (\hat{x}_1, \ldots \hat{x}_d)^T \in \mathbb{R}^d$

We know, $P(Y=y \mid \hat{x}) = \dfrac{P(\hat{x} \mid Y=y)\, P(Y=y)}{P(\hat{x})}$ ← Its sufficient to maximise this.

Naive Bayes (NB)

Assumption: $P(x \mid Y=y) = \prod_{i=1}^{d} P(x_i \mid Y=y)$

∴ we've:

$$\underset{y}{\operatorname{argmax}}\, P(Y=y \mid x) = \underset{y}{\operatorname{argmax}} \left(\prod_{i=1}^{d} P(x_i \mid Y=y)\, P(Y=y) \right)$$

$$= \underset{y}{\operatorname{argmax}} \left\{ \sum \log P(x_i \mid Y=y) + \log P(Y=y) \right\}$$

Observe:

$P(x_i \mid Y=y) = \dfrac{\#\text{ of times } x_i \text{ and } y \text{ happen together}}{\#\text{ of times } Y=y}$ $\qquad P(Y=y) = \dfrac{\#\text{ of times } Y=y}{n}$

Assuming binary for x_i's, $(d \times 1)$ parameter for $x \in \mathbb{R}^d$ — If there are k classes for y, no. of parameters $= (k-1)$ $\Rightarrow$ $d \times (k-1)$ parameters for $P(x_i=1 \mid Y=y)$

But for $P(x \mid Y=y)$ we need $2^{d-1} \times (k-1)$ parameters (why?)

Uses of NB Classifier: Topic Classification

Given a document, we need to find its topic

Treat the document as a "bag of words".

Given sentences, we create a vocabulary of words, and index them arbitrarily (tokenization)

arbitrarily (tokenization)

Article : $x = \{x_1, \dots, x_N\}$ where x_i is the i^{th} word.

$$P(x_i \mid Y=y) = \frac{\#\text{ of times } x_i \text{ appears in doc of class } y + 1}{\sum_{w \in V} (\#\text{ of times } w \text{ appear in } y) + |V|}$$

V = Vocabulary of words

Because if some new word come during test, we don't want the prob. to go to zero

$$P(x \mid Y=y) = \prod_{i=1}^{N} P(x_i \mid Y=y)$$

$P(Y=y)$ = relative fraction of y in all topics.

Instead of looking at single words, we can extend x_i for a group of words

Disadvantages of NB

Consider a true distribution of binary variables X_1 and Y such that :

	$Y=0$	$Y=1$
$P(Y)$	0.8	0.2

$P(Y)$

	$X_1=0$	$X_1=1$
$Y=0$	0.7	0.3
$Y=1$	0.3	0.7

$P(X_1 \mid Y)$

we've: $\hat{y} = \underset{y}{\text{argmax}}\ P(Y=y \mid X_1) = \underset{y}{\text{argmax}}\ P(X_1 \mid Y=y)\, P(Y=y)$

	$P(X_1 \mid Y=0)$	$P(X_1 \mid Y=1)$	$\hat{y}$
$X_1=0$	0.56	0.06	0
$X_1=1$	0.24	0.14	0

Expected error $= P(Y \neq \hat{y})$

$= P(Y \neq \hat{y}, X_1=0) + P(Y \neq \hat{y}, X_1=1)$

$= P(Y=1, X_1=0) + P(Y=1, X_1=1)$

$= 0.06 + 0.14 = \underline{\underline{0.2}}$

Now we've another X_2 identical to X_1 (i.e a copy of X_1)

	$P(X_1X_2\|Y=0)$	$P(X_1X_2\|Y=1)$	$\hat{y}$
$X_1=X_2=0$	$0.7^2\times0.8$	$0.3^2\times0.2$	0
$X_1=X_2=1$	$0.3^2\times0.8$	$(0.7)^2\times0.2$	1
$X_1=0\ X_2=1$	–	–	0
$X_1=1, X_2=0$	–	–	0

Expected error:

$$P(Y\neq\hat{y}) = P(Y=1, X_1=0) + P(Y=0, X_1=1)$$
$$= 0.3$$

Thus bringing in another highly corelated variable increases errors!

Logistic Regression (Classification)

Suppose, we are interested in a Binary Classification.

From regression, $y \simeq w^T x$.

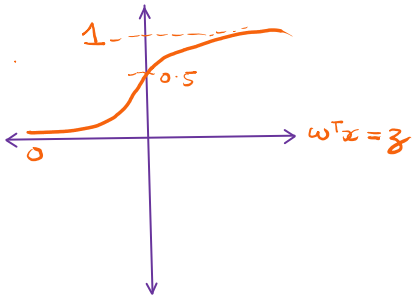
Here if x comes from class 1, $w_1^T x$ will dominate

" " class 2, $w_2^T x$ will dominate.

Logistic fn:-

$$\left[\frac{e^{w_1^T x}}{e^{w_1^T x}+e^{w_2^T x}} \quad \frac{e^{w_2^T x}}{e^{w_1^T x}+e^{w_2^T x}}\right] \rightarrow \frac{1}{1+e^{(w_2-w_1)^T x}} = \frac{1}{1+e^{-w^T x}} = \sigma(w^T x)$$

where $w = w_1 - w_2$

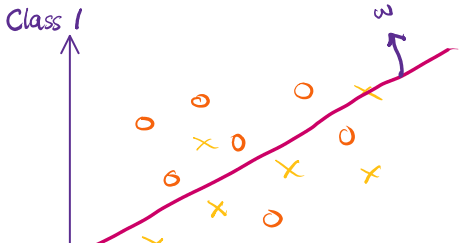


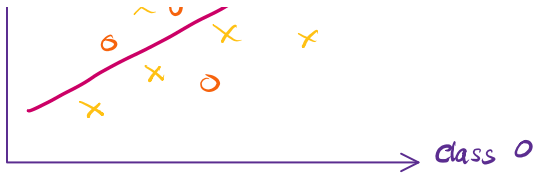
Assumption: $P(Y=1|x,w) = \sigma(w^T x)$

$$\frac{P(Y=1|x,w)}{P(Y=0|x,w)} > 1 \Rightarrow \hat{y}=1$$

otherwise, $\hat{y}=0$

$$\exp(w^T x) > 1 \Rightarrow w^T x > 0$$





Linear Classifier