Step into Machine Learning

1 Predicting the Price of a Property

We talked about a city, named "Birjand". We realized that houses located in the southern areas are more expensive. An efficient way to **represent** where a property is placed, is to use an **ordered pair** that represents its x and y **coordinates**. We considered the south-western corner of the town as the **origin** of the coordinates. However a better way was to use geographic coordinate system, i.e., its **latitude** and **longitude**.

Overall, we **concluded** that these five **factors** are **determinant** of the price of a property:

1. latitude
2. longitude
3. age (The number of years that has passed from building the house)
4. area (**in** square meters)
5. type (i.e. house, apartment, or duplex)

It was revealed that a real estate agent is able to accurately **estimate** the price of a property, if they are given the value of these five **features**. Four of these features are **numerical**[1], but the 'type' is a "..... feature". It should be described by a word. However, sicne we're considering a certain set of types, i.e, 'house', 'apartment', and 'duplex', we can **assign** a number to each of them:

type	value
house	0
apartment	1
duplex	2

Then, we may **come up** with a **feature vector** for each house in Birjand:

$$\vec{f}_{1\times 5} = [\text{lat}, \text{long}, \text{age (in years)}, \text{area (in } m^2), \text{type } (\in \{0, 1, 2\})] \qquad (1)$$

The dimension of feature vectors is 1×5, since we have five features. This is also equivalent to saying that "we have a *five dimensional* ***feature space***".

What we want is a function that maps a feature vector to the price, however, it doesn't seem promising to try to devise an explicit formula for this. The

[1] I used the adjective 'numeral' in the class, which was wrong here.

experts (the real estate agents) are also incapable of telling us how they geuss the prices. What we can do, is to collect labeled data. This means that we can compute the feature vector for properties that we already know their prices. Doing this for a sufficient number of samples, we will have a **dataset** of feature vectors and their correct prices. Formally, a dataset of S samples, will be of the form: $\{(\vec{f_1}, p_1), (\vec{f_2}, p_2), \ldots, (\vec{f_S}, p_S)\}$, where p_i is the **ground truth** for samples and $i = 1, \ldots, S$.

By plotting the data, we realized that a **linear model can fit them**. This was our **hypothesis** for the problem. Relation (2) formulates the concepts:

$$y = \vec{f} \cdot \vec{w}^T + \vec{b} \tag{2}$$

Where $\vec{w}$ is the slope of the line, or **weights**; and $\vec{b}$ resembles the intercept, or off-set.

This is all good, but we don't know what the correct value for $\vec{w}$ is. *Machine learning* is to program the computer to automatically compute the optimum weights, given the dataset. To have an idea how this may be possible, let's call the predicted value for sample i, y_i. Also, assume that we begin by **randomly initialized weights**. So, the error made on sample i, is $e_i = (y_i - p_i)^2$. It is obvious that we desire the error be as small as possible. We also want to know how to determine the weights, so that this happens.

The answer is that we consdier the error as a function of weights. If we expand the error's relation, we will get:

$$e_i(\vec{w}, \vec{b}) = ((\vec{f_i} \cdot \vec{w}^T + \vec{b}) - p_i)^2 \tag{3}$$

This may not be mathematically rigorous, but now, we can think of error as a function of $\vec{w}$ and $\vec{b}$; then we can find for what value of the independent varaibles, the function will be minimized. Derivation is an option for this **optimization** task.

2 Exercise

1. Find other usages for the highlighted words and phrases in technical texts.
2. Find out how to pronounce math expressions like $\int_0^1 f(x).dx$ from Lawrence A Chang's book, *Handbook for Spoken Mathematics*.
3. Read the first chapter of Tom M. Mitchell's *Machine Learning*, New York, McGraw-hill, 1997. (Ask me if you find anything in it that is hard to understand.)