AI Jungle Docs

EsharkyTheGreat

11-05-2024

Contents

1 Probability **1**
- 1.1 Introduction . . . 1
- 1.2 Two Discrete Random Variables . . . 1
- 1.3 Continuous Random Variables . . . 2
- 1.4 Bayes Rule . . . 3
- 1.5 Expectation . . . 4
- 1.6 Variance . . . 4
- 1.7 Covariance . . . 4
- 1.8 Gaussian Distribution . . . 5
  - 1.8.1 Mean . . . 5
  - 1.8.2 Variance . . . 5
  - 1.8.3 Multivariate Gaussian Distribution . . . 6
- 1.9 Maximum Likelihood Principle . . . 6
  - 1.9.1 Maximum Likelihood for Gaussian Distribution . . . 7

2 Statistics **8**
- 2.1 Correlation . . . 8

3 Data Processing **9**
- 3.1 Data Splitting . . . 9
  - 3.1.1 Random Splitting . . . 9
  - 3.1.2 Checksum Splitting . . . 9
  - 3.1.3 Stratified Splitting . . . 10

4 Machine Learning **11**
- 4.1 Linear Regression . . . 11
- 4.2 Logistic Regression . . . 11
- 4.3 Error . . . 11

5 Python Tips and Tricks **12**
- 5.1 Sklearn . . . 12
- 5.2 Matplot . . . 12
- 5.3 Numpy . . . 12
- 5.4 Pandas . . . 12

1 Probability

1.1 Introduction

Probability is a measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur.

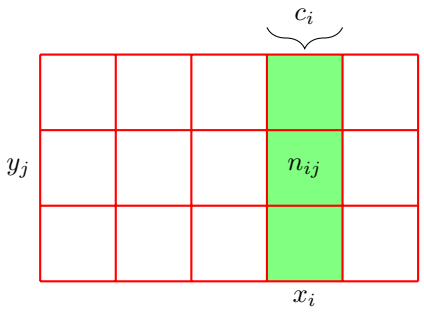
Frequentist Probability fraction of times event occurs in experiment

Bayesian Probability degree of belief in an event

Random Variable X

- Stochastic variable sampled from a set of possible outcomes
- Discrete or continuous E.g Labels like { Heads, Tails } is discrete and gaussian distribution is continuous
- Probability distribution $p(X)$

1.2 Two Discrete Random Variables



x_i = possible values of X
y_j = possible values of Y
n_{ij} = number of times $X = x_i$ and $Y = y_j$
c_i = number of times $X = x_i$
N = total number of samples

2 Random Variables X, Y

$$X = \{x_1, x_2, x_3, x_4, x_5\}$$
$$Y = \{y_1, y_2, y_3\}$$

Joint Probability

$$p(X = x_i, Y = y_j) = n_{ij}/N$$

Marginal Probability

$$\begin{aligned} p(X = x_i) &= c_i/N \\ c_i &= \sum_{j=1}^{3} n_{ij} \\ n_{ij} &= p(X = x_i, Y = y_j) * N \\ p(X = x_i) &= \frac{1}{N} \sum_{j=1}^{3} p(X = x_i, Y = y_j) * N \\ p(X = x_i) &= \sum_{j=1}^{3} p(X = x_i, Y = y_j) \end{aligned}$$

Sum Rule

$$p(X = x_i) = \sum_{Y} p(X = x_i, Y = y_j)$$

Conditional Probability of Y given X

$$\begin{aligned} p(X = x_i) &= \frac{c_i}{N} \\ p(Y = y_j | X = x_i) &= \frac{n_{ij}}{c_i} \\ p(Y = y_j | X = x_i) &= \frac{p(X = x_i, Y = y_i) * N}{p(X = x_i) * N} \\ p(Y = y_j | X = x_i) &= \frac{p(X = x_i, Y = y_j)}{p(X = x_i)} \end{aligned}$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= p(Y = y_j | X = x_i) * p(X = x_i) \\ p(X = x_i, Y = y_j) &= p(X = x_i | Y = y_j) * p(Y = y_j) \end{aligned}$$

1.3 Continuous Random Variables

- Probability of falling in the interval $(x, x + dx)$ is given by $p(x)dx$

- Here $p(x)dx$ is the probability density function over all possible x outputting the probability of its occurence
- Probability over finite interval (a, b) is given by

$$p(x\epsilon(a,b)) = \int_a^b p(x)dx$$

- Positivity

$$p(x) \geq 0$$

- Normalization

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

- Change of variables - $x = g(y)$ probabilities in $(x, x + dx)$ must be transformed to $(y, y + dy)$

$$p_x(x)dx = p_y(y)dy$$
$$p(y) = p(x)|\frac{dx}{dy}|$$

- Sum Rule

$$p(x = x_i) = \int_Y p(x = x_i, y)dy$$

1.4 Bayes Rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- $p(y|x)$ - is the posterior probability of Y=y
- $p(x|y)$ - is the likelihood of X=x given Y=y
- $p(y)$ - is the prior probability of Y=y
- $p(x)$ - is the evidence for X=x

1.5 Expectation

Random Variable $x \epsilon X$ and function $f : X \to R$

$$\begin{aligned}
\mathbb{E}[f] &= \mathbb{E}_{x\sim p(X)}[f(x)] \\
\mathbb{E}[f(x)] &= \sum_x f(x)p(x) \quad \text{Discrete} \\
\mathbb{E}[f(x)] &= \int f(x)p(x)dx \quad \text{Continuous} \\
\mathbb{E}[f(x)+g(x)] &= \mathbb{E}[f(x)] + \mathbb{E}[g(x)] \\
\mathbb{E}[cf(x)] &= c\mathbb{E}[f(x)]
\end{aligned}$$

1.6 Variance

The variance of a random variable is a measure of how much the values of the variable vary as we sample different values of the variable from its probability distribution.
The expected quadratic distance of the random variable from its mean.

$$\begin{aligned}
Var[f] &= \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \\
Var[f] &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\
\mathbb{E}[f(x)] &= constant \\
Var[f] &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 \\
Var[f] &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2
\end{aligned}$$

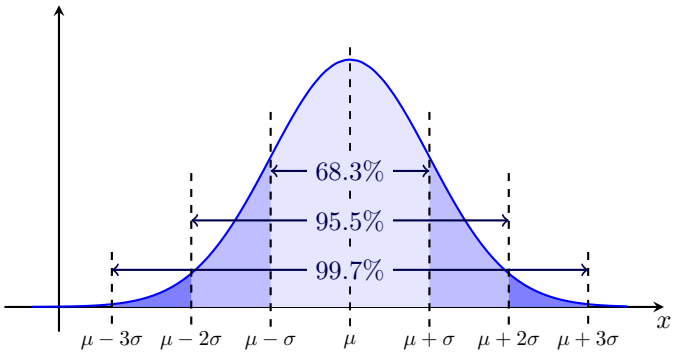
1.7 Covariance

Measure of how much two random variables change together

$$Cov[x,y] = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \tag{1}$$
$$Cov[x,y] = \mathbb{E}[xy - x\mathbb{E}[y] - y\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y]] \tag{2}$$
$$Cov[x,y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] \tag{3}$$
$$\tag{4}$$

Independent Random Variables $p(x,y) = p(x)p(y)$
Covariance doesn't imply that variables are independant

1.8 Gaussian Distribution



$$
\begin{aligned}
p(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp - \frac{(x-\mu)^2}{2\sigma^2} \\
\mathcal{N}(x|\mu,\sigma^2) &= \text{Normal Distribution} \\
\mu &= \text{mean} \\
\sigma^2 &= \text{variance}
\end{aligned}
$$

Since this is a probability distribution it should be normalized that is with respect to the function $f(x) = 1$ the expected value should be 1 or the area under the curve of this distribution should be one

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2}) dx = 1$$

1.8.1 Mean

$$
\begin{aligned}
\mathbb{E}[x] &= \int_{-\infty}^{\infty} x\mathcal{N}(x|\mu,\sigma^2)dx \\
\mathbb{E}[x] &= \mu
\end{aligned}
$$

1.8.2 Variance

$$
\begin{aligned}
Var[x] &= \mathbb{E}[(x - \mathbb{E}[x])^2] \\
Var[x] &= \int_{-\infty}^{\infty} (x-\mu)^2 \mathcal{N}(x|\mu,\sigma^2)dx \\
Var[x] &= \sigma^2
\end{aligned}
$$

1.8.3 Multivariate Gaussian Distribution

x is a D-dimensional vector $x = (x_1, x_2, x_3, ..., x_D)^T$

$$
\begin{aligned}
p(x) &= \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \\
\Sigma &= \text{Covariance Matrix} \\
\Sigma &= cov[x, x] \\
\mathbb{E}[x] &= \int x\mathcal{N}(x|\mu, \Sigma)dx \\
&= \mu
\end{aligned}
$$

1.9 Maximum Likelihood Principle

- Given a dataset $D = (x_1, x_2, x_3, ..., x_N)$ and a model $p(x|\theta)$ the likelihood of the model is given by $p(D|\theta)$ where θ is the parameter of the model
- The maximum likelihood estimate of θ_{ml} is the value that maximizes the likelihood of the model given the data

$$\theta_{ml} = \arg\max_{\theta} p(D|\theta)$$

- We assume $(x_1, x_2, ..., x_N)$ are independent and identically distributed

$$
\begin{aligned}
p(D|\theta) &= \prod_{n=1}^{N} p(x_n|\theta) \\
\theta_{ml} &= \arg\max_{\theta} \prod_{n=1}^{N} p(x_n|\theta)
\end{aligned}
$$

- Instead of likelihood we maximize log likelihood as probabilities multiplied together will result in a very small number, it can be proven that taking log doesnt affect the result

$$
\begin{aligned}
\theta_{ml} &= \arg\max_{\theta}(log(\prod_{n=1}^{N} p(x_n|\theta))) \\
\theta_{ml} &= \arg\max_{\theta} \sum_{n=1}^{N} log(p(x_n|\theta))
\end{aligned}
$$

1.9.1 Maximum Likelihood for Gaussian Distribution

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$
$$p(D|\mu,\sigma^2) = \prod_{n=1}^{N} p(x_n|\mu,\sigma^2)$$
$$log(p(D|\mu,\sigma^2)) = \sum_{n=1}^{N} log(p(x_n|\mu,\sigma^2))$$

For finding the maximum likelihood estimate we take the derivative of the log likelihood with respect to the parameters and set them to zero

$$\frac{\partial}{\partial\mu} log(p(D|\mu,\sigma^2)) = 0$$
$$\frac{\partial}{\partial\sigma^2} log(p(D|\mu,\sigma^2)) = 0$$

Solving this we get the maximum likelihood estimate of the parameters

$$\mu_{ml} = \frac{1}{N}\sum_{n=1}^{N} x_n$$
$$\sigma^2_{ml} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{ml})^2$$

2 Statistics

2.1 Correlation

Pearson's Correlation Coefficient The correlation coefficient ranges from –1 to 1. When it is close to 1, it means that there is a strong positive correlation; for example, the median house value tends to go up when the median income goes up. When the coefficient is close to –1, it means that there is a strong negative correlation; you can see a small negative correlation between the latitude and the median house value (i.e., prices have a slight tendency to go down when you go north). Finally, coefficients close to zero mean that there is no linear correlation

3 Data Processing

3.1 Data Splitting

It is important that we immediately split the data into training and testing sets so that there is no data snooping bias (Our brain always looks at pattern and is good at overfitting therefore when we look at the whole data we might apply techniques that overfit the data and the model becomes unusable).

3.1.1 Random Splitting

We take a random seed and split the data into training and testing sets. But the problem is that the data might not be evenly distributed and if we add more data the split might change due to this some data in the original training set will come in the testing set causing bias.

```
import numpy as np

def split_train_test(data, test_ratio):
    np.random.seed(42)
    shuffled_indices = np.random.permutation(len(data))
    test_set_size = int(len(data) * test_ratio)
    test_indices = shuffled_indices[:test_set_size]
    train_indices = shuffled_indices[test_set_size:]
    return data.iloc[train_indices], data.iloc[test_indices]
```

3.1.2 Checksum Splitting

We calculate the hash of the data and split the data based on the hash. This way the data will always be split the same way and we can add new data without worrying about the split changing.

```
from zlib import crc32

    def test_set_check(identifier, test_ratio):
        return crc32(np.int64(identifier)) & 0xffffffff < test_ratio * 2**32
    def split_train_test_by_id(data, test_ratio, id_column):
        ids = data[id_column]
        in_test_set = ids.apply(lambda id_: test_set_check(id_, test_ratio))
        return data.loc[~in_test_set], data.loc[in_test_set]
```

3.1.3 Stratified Splitting

When we want the same distribution of categorical data in the main dataset to be present in both the training and testing data set because of the importance of the distribution to the model we use Stratified Splitting

```
from sklearn.model_selection import StratifiedShuffleSplit

split = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
for train_index, test_index in split.split(housing, housing["income_cat"]):
    strat_train_set = housing.loc[train_index]
    strat_test_set = housing.loc[test_index]
```

4 Machine Learning

4.1 Linear Regression

4.2 Logistic Regression

4.3 Error

Root Mean Square Error (RMSE) - L2 Norm

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(x_i) - \hat{y_i})^2} \tag{5}$$

$$\begin{aligned} where, \quad h(x_i) &= \text{predicted value} \\ h &= \text{hypothesis function} \\ \hat{y_i} &= \text{actual value} \end{aligned}$$

This is useful when you want to know how far off your predictions are from the actual values. Cases where the model is far from the correct value is treated more harshly then the cases that are nearby but not exactly correct.

Mean Absolute Error (MAE) - L1 Norm

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|h(x_i) - \hat{y_i}| \tag{6}$$

$$\begin{aligned} where, \quad h(x_i) &= \text{predicted value} \\ h &= \text{hypothesis function} \\ \hat{y_i} &= \text{actual value} \end{aligned}$$

We mostly use RMSE but when there are too many outlier cases that we can't have the model be harsh on all of them we use MAE that averages out the error.

5 Python Tips and Tricks

5.1 Sklearn

5.2 Matplot

5.3 Numpy

5.4 Pandas