Part I

Math

Chapter 1.

Probability

1.1. standart deviation

standart deviation

$$\sigma = \sqrt{Var} = \sqrt{\sigma^2}$$

Var - variance (дисперсия)
1σ - 68% данных для нормального распределения, 95% - для двух.

1.2. Variance

variance

$$Var(X) = E[(X - E[X])^2]$$

1.3. Bayes rule

Joint probability, $P(A, B) = P(A \cap B)$ - это вероятность того, что произошло событие A и B одновременно. $P(A, B) = P(A|B)P(B)$, если события A и B независимы то $P(A|B) = P(A)$ следовательно получаем такую формулу: $P(A, B) = P(A)P(B)$. $P(A, B) = P(B, A)$, отсюда можно получить **Bayes Theorem**:

$$P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

1.4. Prior probability and Posterior probability

Исходя из теоремы байеса:

$$posterior \propto prior \times likelyhood$$

Posterior probability вероятность события после какого то измерения. $P(A|B)$, если рассматривать в терминах машинного обучения $P(H|D)$, где H - гипотеза или модель, D - данные.
Prior probability вероятность распределения гипотезы.

Пример

$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$, H - имеется заболевание (гипотеза, модель), E - тест на заболевание положителен (результаты/данные/событие).
$P(H)$ - **Prior probability** того, что гипотеза правдоподобна (в данном примере какова вероятность того что есть заболевание до того как был проведен тест, то есть насколько распространено заболевание).
$P(E|H)$ - насколько вероятно что тест будет положителен если действительно имеется заболевание.
$P(E)$ - просто вероятность того, что тест положительный. Это комбинация $P(H)P(E|H)$ и $P(!H)P(E|!H)$

$$P(E) = P(H)P(E|H) + P(!H)P(E|!H)$$

Prior probability (что гипотеза верна) наиболее сложно определить на практике.
Теперь посчитаем пример: Пришел положительный тест на заболевание, заболевание распространено у 10% населения. Точность теста равна 95%. Какова вероятность того что заболевание действительно есть если пришли положительные результаты теста.

$$P(H|E) = \frac{P(E|H)P(H)}{P(H)P(E|H) + P(!H)P(E|!H)}$$

$$P(H|E) = \frac{0.95 * 0.1}{0.95 * 0.1 + 0.9 * 0.05} = 0.67$$

В спаме:

$$P(spam|word) = \frac{p(word|spam)P(spam)}{P(word)}$$

Conjugate prior In Bayesian probability theory, if the posterior distributions p(θ | x) are in the same probability distribution family as the prior probability distribution p(θ), the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

1.5. Normal Distribution

Normal Distribution распределение вероятностей, которое в одномерном случае задаётся функцией плотности вероятности, совпадающей с функцией Гаусса

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-E[x])^2}{2\sigma^2}}$$

1.6. Beta Distribution

Beta Distribution is a family of continuous probability distributions defined on the interval [0, 1] parametrized by two positive shape parameters, denoted by α and β. In Bayesian inference, the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions.

$$P(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

The **beta function**, B, is a normalization constant to ensure that the total probability is 1. The beta function, also called the **Euler integral** of the first kind.

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}$$

1.7. Cauchy distribution

Cauchy distribution или (Распределение Коши). Случайная величина, имеющая распределение Коши, является стандартным примером величины, не имеющей математического ожидания и дисперсии.

$$F_X(x) = \frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right) + \frac{1}{2}$$

1.8. Confidence intervals

Confidence intervals There is a 95% chance that p(x) is within $2\sigma_{\hat{p}}$ of $\hat{p}(x)$

1.9. Hypothesis testing

Null-Hypothesis, Нулевая гипотеза — принимаемое по умолчанию предположение о том, что не существует связи между двумя наблюдаемыми событиями, феноменами. Так, нулевая гипотеза считается верной до того момента, пока нельзя доказать обратное.

t-statistcs :
The T Statistic is used in a T test when you are deciding if you should support or reject the null hypothesis. It's very similar to a Z-score and you use it in the same way: find a cut off point, find your t score, and compare the two. You use the t statistic when you have a small sample size, or if you don't know the population standard deviation.

p-value :
сначала устанавливаем significance level (threshold) - α

Например возьмем сэмпл и получим, что среднее значение = 25.

$$P(\mu_{sample} \geq 25 | H_0 == True)$$

$$p - value < \alpha \Longrightarrow \text{reject} H_0 \text{accept} H_a$$

$$p - value \geq \alpha \Longrightarrow \text{do not reject} H_0$$

When you run a hypothesis test, you use the T statistic with a p value. The p-value tells you what the odds are that your results could have happened by chance. Let's say you and a group of friends score an average of 205 on a bowling game. You know the average bowler scores 79.7. Should you and your friends consider professional bowling? Or are those scores a fluke? Finding the t statistic and the probability value will give you a good idea. More technically, finding those values will give you evidence of a significant difference between your team's mean and the population mean (i.e. everyone).

The greater the T, the more evidence you have that your team's scores are significantly different from average. A smaller T value is evidence that your team's score is not significantly different from average. It's pretty obvious that your team's score (205) is significantly different from 79.7, so you'd want to take a look at the probability value. If the p-value is larger than 5%, the odds are your team getting those scores are due to chance. Very small (under 5%), you're onto something: think about going professional.

The **Z-score** allows you to decide if your sample is different from the population mean. In order to use z, you must know four things:

1) The population mean. 2) The population standard deviation. 3) The sample mean. 4) The sample size.

Chapter 2.

Linear algebra

2.1. Identity and Inverse Matrices

We denote the identity matrix that preserves n-dimensional vectors as I_n:

$$\mathbf{I}_n\mathbf{x} = \mathbf{x}$$

The structure of the identity matrix is simple: all of the entries along the main diagonal are 1, while all of the other entries are zero. Единичная матрица квадратная.

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Таким образом мы можем определить обратную матрицу как: $\mathbf{A}^{-1}$, для нее справедливо:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$$

Part II

Machine learning

Chapter 3.

ML algorithms

3.1. SVM

SVM - Support vector machine

Chapter 4.

Loss Functions

4.1. MSE, OLS

Все это одно и то же по сути, **RSS** - residual sum of squares, **OLS** - ordinary least squares, **LS** - least squares, **MSE** - mean squared error, **SE** - squared error. В разных источниках можно встретить разные названия. Суть у этого всего одна: квадратичное отклонение. Можно запутаться конечно, но к этому быстро привыкаешь.

Стоит отметить, что MSE это средне-квадратичное отклонение, некое среднее значение ошибки для всего тренировочного набора данных. На практике обычно MSE и используется. Формула особо ничем не отличается:

$$MSE(\beta) = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

N - размер датасета, $\hat{y}_i$ - предсказание модели для y_i.

Chapter 5.

Linear regression

Обычная линейная модель:$\hat{y}(x) = f_\theta = \sum_{i=1}^{n} x_i\theta_i + \theta_0$. В векторном (матричном если тренируем сразу на батче) виде: $\hat{Y}(X) = \mathbf{X}^T\theta$.
При этом смещение (bias) θ_0 поместим в общий вектор, $x_0 = 1$.Используем функцию ошибки **RSS** (residaul sum of squares). Почти тоже самое, что и **MSE** (Смотреть в разделе section 4.1).

$$L = RSS = \sum_{i}^{N} (y_i - \hat{y}_i)^2$$

Или в векторном виде (далее в этом разделе все будет в вектрном виде):

$$L(\mathbf{X}) = (\mathbf{Y} - \hat{\mathbf{Y}})^2 = (\mathbf{Y} - \mathbf{X}^T\theta)^2$$

Наша цель найти параметры θ, для этого возьмем частную производную от функции ошибки L по θ и приравняем ее к нулю.

$$\frac{\delta L}{\delta \theta} = \frac{1}{2}(\mathbf{Y} - \mathbf{X}^T\theta)\mathbf{X} = 0$$

Отсуда можно найти θ:

$$(\mathbf{Y} - \mathbf{X}^T\theta)\mathbf{X} = \mathbf{Y}^T\mathbf{X} - \mathbf{X}^T\mathbf{X}\theta = 0$$

$$\theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}^T\mathbf{X}$$

Это прямой способ нахождения параметров, минусы заключаются в том, что для этого надо находить обратную матрицу $(\mathbf{X}^T\mathbf{X})^{-1}$, а она не всегда существует (плюс она должна быть квадратной, смотреть section 2.1). Такой способ называется решение на прямую или через **The Normal Equation** Так же: размерность обратной матрицы $n \times n$ The computational complexity of inverting such a matrix is typically about $O(n^2.4)$ to $O(n^3)$ (depending on the implementation). On the positive side, this equation is linear with regards to the number of instances in the training set (it

is O(m)), so it handles large training sets efficiently, provided they can fit in memory. Использование итерационного метода (**Gradient Descent** или **Batch Gradient Descent**): Далее будет говорится о Batch Gradient Descent. Поэтому будем использовать **MSE** в качетсве ошибки:

$$L = MSE = \frac{1}{m}\sum_{i}^{N}(y_i - \hat{y}_i)^2$$

где m-размер батча.

$$\frac{\delta MSE}{\delta \theta_j} = \frac{1}{m}\sum_{j}^{m}(\mathbf{y}_j - \mathbf{x}_j^T\theta)\mathbf{x}_j$$

В данном случае к нулю приравнивать не надо, наща цель найти градиент функции стоимости (или функции ошибки, в данном случае MSE).

$$\nabla\theta = \frac{1}{m}(\mathbf{Y} - \mathbf{X}^T\theta)\mathbf{X}$$

Отсюда получаем значения параметров на следующем шаге:

$$\theta_{(nextstep)} = \theta - \eta\nabla\theta$$

r-squared - насколько хорошо данные обьясняются моделью.

5.1. Logistic regression

Logistic regression предсказывает вероятность $[0, 1]$

Chapter 6.

Decision Tree

Chapter 7.

Bayesian Learning

Suppose that you are allowed to flip the coin 10 times in order to determine the fairness of the coin. Your observations from the experiment will fall under one of the following cases:

- Case 1: observing 5 heads and 5 tails.
- Case 2: observing h heads and 10-h tails, where $h \neq 10 - h$.

If case 1 is observed, you are now more certain that the coin is a fair coin, and you will decide that the probability of observing heads is with more confidence. If case 2 is observed you can either:

1) Neglect your prior beliefs since now you have new data, decide the probability of observing heads is h/10 by solely depending on recent observations.

2) Adjust your belief accordingly to the value of that you have just observed, and decide the probability of observing heads using your recent observations.

The first method suggests that we use the frequentist method, where we omit our beliefs when making decisions. However, the second method seems to be more convenient because 10 coins are insu"cient to determine the fairness of a coin. Therefore, we can make better decisions by combining our recent observations and beliefs that we have gained through our past experiences. It is this thinking model which uses our most recent observations together with our beliefs or inclination for critical thinking that is known as Bayesian thinking.
Moreover, assume that your friend allows you to conduct another coin flips. Then we can use these new observations to further update our beliefs. As we gain more data, we can incrementally update our beliefs increasing the certainty of our conclusions. This is known as incremental learning, where you update your knowledge incrementally with new evidence.
Bayesian learning comes into play on such occasions, where we are unable to use frequentist statistics due to the drawbacks that we have discussed above. We can use Bayesian learning to address all these drawbacks and even with additional capabilities

(such as incremental updates of the posterior) when testing a hypothesis to estimate unknown parameters of a machine learning models. Bayesian learning uses Bayes' theorem to determine the conditional probability of a hypotheses given some evidence or observations.

7.1. Maximum a Posteriori (MAP)

We can use **MAP** to determine the valid hypothesis from a set of hypotheses. According to MAP, the hypothesis that has the maximum posterior probability is considered as the valid hypothesis. Therefore, we can express the hypothesis θ that is concluded using MAP as follows:

$$\theta = argmax_\theta P(\theta_i|X)$$

7.2. Bayesian Learning

Предпологаем, что: The prior, likelihood, and posterior are continuous random variables that are described using probability density functions. Let us now attempt to determine the probability density functions for each random variable in order to describe their probability distributions.
возьмем: Binomial Likelihood
Beta Prior Distribution
Reasons for choosing the beta distribution as the prior as follows:

- If the posterior distribution has the same family as the prior distribution then those distributions are called as conjugate distributions, and the prior is called the conjugate prior. Beta prior acts as a conjugate prior to Binomial likelihood. If we use a Beta distribution to represent our belief, then the resulting posterior distribution will also be a beta distribution. When we observe new data, then this posterior can be used as the new prior to compute the new posterior. Therefore, we can incrementally update the prior whenever new data is available avoiding many complex computations from Bayes' theorem — the posterior can be derived by altering the shape parameters of the conjugate priors accordingly.

- Beta distribution has a normalizing constant, thus it is always distributed between and . Therefore we are not required to compute the denominator of the Bayes' theorem to normalize the posterior probability distribution — Beta distribution can be directly used as a probability density function of θ(recall that θ is also a probability and therefore it takes values between 0 and 1).

- When we are unable to decide the prior distribution due to lack of past experience, we can use the uninformative prior with minimal influence on the posterior. An uninformative prior can be generated by setting the shape parameters of Beta distribution $\alpha = \beta = 1$

Если бета распределение не информативное, то постериорное распределение более похоже на likelihood.

Bayesian learning is capable of incrementally updating the posterior distribution whenever new evidence is made available while improving the confidence of the estimated posteriors with each update.

7.3. Linear Regression with BL

Для решения задачи регресии можно использовать least squared error or using the maximum likelihood, which is categorized as frequentist methods. According to the frequentist method, we can determine a single value per each parameter. Moreover, the frequentist method gives exact point estimations to the unknown model parameters (W) and does not show the variation of those model parameters.

Confidence interval guarantees with a certain confidence that the estimated value lies within a certain interval, whereas concepts of uncertainty in Bayesian learning measures the confidence of the each value from the estimated posterior distributions.

Модель:

$$y = wx + b$$

$$\mu = y = wx + b$$

$$y \sim \mathcal{N}(wx + b, \sigma^2)$$

Therefore, we have to determine the parameters w, b, σ^2, and using Bayesian inference given the X and Y.

Используя теорему байеса, получаем:

$$\underbrace{P(w, b, \sigma^2|Y, X)}_{posterior} = \frac{\underbrace{P(Y|w, b, \sigma^2, X)}_{likelihood}\underbrace{P(w, b, \sigma^2)}_{prior}}{\underbrace{P(Y|X)}_{evidence}}$$

Из данной формулы можно убрать evidence так как он не зависит от параметров модели:

$$P(w, b, \sigma^2|Y, X) \propto P(Y|wX + b, \sigma^2)P(w, b, \sigma^2)$$

Учитвая, что переменные независимы формулу можно переписать:

$$P(w, b, \sigma^2|Y, X) \propto \prod_i^N [P(y_i|wx_i + b, \sigma^2)]P(w)P(b)P(\sigma^2)$$

Можно определить априорные распределения как нормальные или распределение Коши для дисперсии.

The process of learning hidden variables is called **Bayesian inference**.

Typically, classical machine learning models are less useful in the absence of su!cient data to train those models. However, a Bayesian model can still be useful with less data owing to its capability to attach an uncertainty value with each prediction.

7.4. MCMC sampling

MCMC sampling or Markov Chain Monte Carlo sampling один из методов нахождения скрытых параметров для байесовского обучения или **Bayesian inference**.

Chapter 8.

Bayesian Optimization

Bayesian optimization is an approach to optimizing objective functions that take a long time (minutes or hours) to evaluate. BayesOpt is designed for black-box derivative free global optimization.

Chapter 9.

Regularizing neural networks, optimizing models

Методы поиска оптимальной модели, подобор гиперпараметров:

- Random: Try random configurations of layers and nodes per layer.
- **Grid Search** Try a systematic search across the number of layers and nodes per layer.
- Heuristic: Try a directed search across configurations such as a genetic algorithm or Bayesian optimization.
- Exhaustive: Try all combinations of layers and the number of nodes; it might be feasible for small networks and datasets.

9.1. Bayesian Hyperparameter Optimization

Chapter 10.

A/B testing

The humble A/B test (also known as a randomised controlled trial, or RCT, in the other sciences) is a powerful tool for product development.

ROI (return on investment)

Конверсия

Конверсия вычисляется как доля от общего числа посетителей, совершивших какое-либо действие. Действием может быть заполнение формы на посадочной странице, совершение покупки в интернет-магазине, регистрация, подписка на новости, клик на ссылку или блок.

Экономические метрики

Как правило, эти метрики применимы для интернет-магазинов: величина среднего чека, объем выручки, отнесенный на число посетителей интернет-магазина.

Поведенческие факторы
9035389769

К поведенческим факторам относят оценку заинтересованности посетителей в ресурсе. Ключевыми метриками являются: глубина просмотра страниц — число просмотренных страниц, отнесенное к числу посетителей на сайте, средняя продолжительность сессии, показатель отказов — доля пользователей, покинувших сайт сразу после первого захода, коэффициент удержания (можно считать, как 1 минус % новых пользователей).

Chapter 11.

Time Series Analysis

Time Series Components
A useful abstraction for selecting forecasting methods is to break a time series down into systematic and unsystematic components.

- **Systematic**: Components of the time series that have consistency or recurrence and can be described and modeled.
- **Non-Systematic**: Components of the time series that cannot be directly modeled.

A given time series is thought to consist of three systematic components including level, trend, seasonality, and one non-systematic component called noise.

- **Level**: The average value in the series.
- **Trend**: The increasing or decreasing value in the series.
- **Seasonality**: The repeating short-term cycle in the series.
- **Noise**: The random variation in the series.

When a time series is stationary, it can be easier to model. Statistical modeling methods assume or require the time series to be stationary.
There are multiple tests that can be used to check stationarity.

- **ADF** (Augmented Dicky Fuller Test)
- **KPSS**
- **PP** (Phillips-Perron test)

KPSS test:
Null Hypothesis: the process is trend-stationary
Alternative Hypothesis: the process has a unit root (this is how the authors of the test

defined the alternative in their original 1992 paper)
ADF test:
Null Hypothesis: the process has a unit-root (”difference stationary”)
Alternative Hypothesis: the process has no unit root. It can mean either that the process is stationary, or trend stationary, depending on which version of the ADF test is used

A Random Walk - временной ряд который выглядит как случайный

In probability theory and statistics, a **unit root** is a feature of some stochastic processes (such as random walks) that can cause problems in statistical inference involving time series models. A linear stochastic process has a unit root, if 1 is a root of the process's characteristic equation. Such a process is non-stationary but does not always have a trend.

autocorrelation (**ACF**)
partial autocorrelation (**PACF**). The PACF is partial correlation between residuals, controlling for shorter lags.

11.1. AR, MA and ARIMA

MA - Next value in the series is a function of the average of the previous n number of values **AR** - The errors(difference in mean) of the next value is a function of the errors in the previous n number of values **ARMA** - a mixture of both.

MA(q) - moving average of q times back.

$$X_t = Z_t + ... + \theta_{t-q} Z_{t-q}$$

where Z_t - noise.
ARIMA
difference transform - for remove trend.
log-return - log of diff.
Akaike Information Criterion and Model Quality **AIC**

Chapter 12.

Additional Info

CHAID - Chi-square automatic interaction detection
Contingency table or Two-way table are used in statistics to summarize the relationship between several categorical variables.

Part III

Programming

Chapter 13.

Alghoritms

13.1. Metropolis–Hastings algorithm

Metropolis–Hastings algorithm или Алгоритм Метрополиса — Гастингса

In statistics and statistical physics, the Metropolis–Hastings algorithm is a Markov chain Monte Carlo (**MCMC**) method for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult.

Index