Slide 1

Good afternoon, dear students. Welcome to the sixth machine learning lesson. My name is Aleksei Platonov and today we’ll talk about another important concept of machine learning, namely about the using of probability theory in solving machine learning problems. Let me remind you that during the course we touched on several basic approaches to solving the problem of machine learning. The first approach is based on the principle of searching for similar objects, it can be formulated as “for unknown X, find the objects most similar to it from the training set and choose the correct answer among them”. The second approach was different in that we applied the rules of logical inference and solved the problem of machine learning as a series of logical conclusions. The third approach is based on the idea of ​​searching for separating surfaces in multidimensional vector spaces. It turned out that this idea is the basis of neural networks that seek to simulate the work of the human brain. So, as I said, today we’ll talk about the fourth approach.

Slide 2

Here is the plan for today's lesson. We start by recalling the basic concepts of probability theory. The use of probability theory in machine learning is based on conditional probabilities and Bayesian inference. We will talk about these concepts. Finally, we formulate a vision of machine learning as problems of probability theory and, based on this, consider two methods for solving problems. Finally, we will talk about the most important principle of probability theory and machine learning, the principle of maximum likelihood, and see how gradient descent, the k-nearest-neighbor method, and probability models are related.

Slide 3

So let's start with the basics. As I said, we will start with the basic concepts of probability theory. Probability theory is based on several concepts - outcomes, events, experiments, probability distribution function. An experiment in probability theory is any sequence of actions that can have multiple outcomes. For example, an experiment might be tossing a coin or responding to a series of machine learning tests. By outcome is meant the result of an experiment. It is assumed that the experiment can have several outcomes, and in the general case, it is impossible to say in advance exactly what kind of outcome the experiment will have before the experiment. That is, there is some uncertainty in the experiment, and it is precisely it that is modeled by the probability distribution function. This function applies to outcomes or events and takes values ​​from zero to one. Moreover, the sum of all its values ​​on all outcomes is equal to one, that is, the probability that the experiment will end in some outcome is equal to one. The larger the value of this function, the more likely the outcome for which this value was calculated. Finally, an event is a collection of outcomes. For example, in an archery experiment, there are ten outcomes. Each outcome in this experiment is a hit from one to ten in the corresponding part of the target. The event will be hitting the center of the target and gaining nine or ten points.

Slide 4

How is the probability distribution function defined? If we are dealing with a discrete and finite set of events, then the simplest way is to build a histogram. Many of you know what a histogram is. On the x-axis on the histogram are all possible outcomes that can occur in the experiment, and the probability of this event is plotted on the y-axis. A histogram is also a great way to study data in a machine learning task, because with it you can see how the data is distributed in the sample.

Slide 5

If the experiment can end with an outcome that takes on a value from a continuous set of outcomes, it is convenient to use the probability density function. This function is designed in such a way that the area under its curve, limited by the range of values of the event, is equal to the probability of this event. Moreover, the area under the entire curve in the entire range of outcomes is equal to one.

Slide 6

The next two important concepts of probability theory are dependent events and conditional probabilities. Two or more events are called independent if their probabilities are independent of each other. Then the probability that two events will occur in the experiment is equal to their product. An example of two independent events is the successive coin flip. The second coin flip does not depend on the previous one. Moreover, if the event depends on another event, then the product of their probabilities will no longer be equal to the probability that these two events will occur. If event X2 depends on X1, then the concept of conditional probability is introduced, that is, the probability that event X2 will occur if event X1 occurs. Their product will be equal to the probability of these two events. For example, the events “meet the seller of lottery tickets” and “win the lottery” in one context will be dependent events.

Slide 7

The concept of conditional probability leads us to a very important law, namely the Bayes theorem. Let's look at this theorem a little closer. According to the classical probabilities’ theory, the probability of events A and B does not depend on the sequence in which we consider them. Those the probability of events A and B is equal to the probability of events B and A. Then each of two such events can be described using the conditional probability formula. Moreover, these probabilities are equal, which means that you can come to an expression in the center of the slide. This expression connects the conditional probabilities of two dependent events. The expression in the denominator is called the total probability of event B, and the expression in the numerator determines the probability that event B will occur provided that event A. occurs. Using the formula for the full probability presented below, you can get the final expression of Bayes' theorem, which can be seen at the bottom of the slide. In this formulation, Aj - is called a hypothesis, and B is an observable event that can occur according to one of the presented hypotheses. Bayes theorem allows us to determine the probabilities of a particular hypothesis, thus determining the possible causes of the observed event.

Slide 8

Interestingly, in this formulation, Bayes' theorem allows us to perform some reasoning. Namely, with the help of Bayes' theorem, we can try to find the causes of the observed phenomenon, determine which cause is the most probable, and which is less likely. This introduces several useful concepts. Namely, the concept of a priori information is introduced. This is our knowledge of the subject area with the help of which we evaluate the probabilities of outcomes and hypotheses before even conducting an experiment. Finally, the concept of posterior information is introduced. This is information at the moment when the expected event occurred and we evaluate the probability of a particular hypothesis.

Slide 9

All this allows us to proceed to the consideration of the problem of machine learning from the point of view of probability theory. After all, we have already considered the approach using reasoning based on logical rules. Let us once again remind ourselves of the formulation of the machine learning problem. The goal of the machine learning task is to determine a decision function or algorithm a, which matches the answers from the set Y to objects from the set X. If there is a training sample that consists of pairs of objects and the correct answers, the machine learning task is called supervised learning. And now let's imagine that some unknown probabilistic law is at the heart of the problem being solved. That is, for all kinds of pairs of objects and answers, some unknown probability distribution function is given, which determines the observed training sample. If at the same time the training sample was formed from pairs of objects and answers independently, then we can try to estimate the parameters of the unknown probability distribution function and thus solve the problem of machine learning.

Slide 10

Then we can connect the concepts of Bayesian inference and machine learning. It is important to note that in the framework of this lecture we will consider only classification problems. A priori information is our knowledge about how objects from the training set are distributed and how answers are distributed. At the same time, which is important, in this context we consider objects and responses independently. In this model, through conditional probability, we connect the object of the training sample X and the answer for this object. The likelihood function is the probability distribution function of vector descriptions of objects for some distinguished class y from the set of answers. The posterior probability of class y when observing the vector X is our desired probability distribution function. The principle of maximum posterior information tells us the following - "the classification of object X results in a label of class Y that has the highest estimate of posterior probability." But how can one determine this posterior probability? Then Bayes' theorem will help us! We can derive the value of posterior information using it. But why is there no p(x) in the expression of the principle of maximum a posteriori information? We can remove this value from under the denominator, since it does not depend on Y, which means that the maximum value will be determined only by the numerator, as this principle speaks of.

Slide 11

But the probabilistic statement of the problem of machine learning is not just interesting in itself. Such a statement of the problem allows us to create a classifier that minimizes the likelihood of making mistakes with the available data for training. Let's introduce some new objects to consider. Ay is the set of all objects from X that will be classified as ‘y’ using a classifier based on the principle of maximum posterior probability. Then the error of such a classifier is all that set of objects from the set Ay, which actually should have the label ‘y’’. Then the probability of such an error can be defined as the sum or integral over the entire set Ay of the probability distribution function on the set of objects and answers with a fixed answer ‘y’’. In other words, we consider the probability with which the classifier maps the answer y to the object X, while the correct answer was y ’.

Slide 12

If we introduce the cost of the error, then we can calculate the expected loss of the classifier, which is equal to the weighted sum of all possible classification errors. If the cost of errors is symmetric for all types of errors, then there is a theorem that says that a classifier using the principle of maximum posterior probability and estimating the cost of errors will have the least chance of error in its work in comparison with other classifiers. Thus, and this is very important, the basis for the probabilistic formulation of the classification problem is the idea of ​​minimizing the probability of error.

Slide 13

Thus, the task of developing a classifier based on probability theory can be represented as two separate tasks. The first task is to evaluate the posterior probabilities of classes and likelihood functions for objects of the space of their vectors based on the training set. The second task is to use this information to obtain the optimal classifier. But we have already solved the second problem! You need to use the Bayesian classifier. However, some correction needs to be made. The fact is that the theorem works in the case when the likelihood and a posteriori information exactly describe the objects of the training sample. But, since we are dealing with a finite training set, we cannot create models that will ideally describe the distribution of objects and classes. Thus, the conditions of the theorem are violated and we get a good classifier, but with errors.

Slide 14

But how to solve the first problem? Let me remind you that we need to evaluate two components: a priori information about the classes and the likelihood of objects with respect to the classes of the problem. To evaluate a priori information, there is nothing easier than using a statistical estimate of the probability of a class in a sample. You can see the formula for this value on the slide. This is simply the frequency with which the selected class y occurs in the training set. But how to describe the likelihood? Let's temporarily move on to the one-dimensional problem, and only for one single class y. Indeed, the likelihood assessment in each class is carried out independently, which means that this can be considered as a separate task.

Slide 15

Moreover, the use of a one-dimensional likelihood estimate is also useful even for a problem with a multidimensional space of object descriptions. The fact is that if we assume that all the features in the description of an object are independent and do not correlate, then a complex estimate of the probability of a multidimensional vector turns into an estimate of the product of the probabilities of individual features. If we take a logarithm of this expression, then it turns into a sum and the final classifier formula is presented at the bottom of the slide. The classifier that uses this assumption is called the naive Bayes classifier. Indeed, he uses a rather strong assumption about the independence of individual features in the description of the object, which in reality rarely happens.

Slide 16

So how can we measure likelihood? In general, there are three main approaches presented on the slide. The first approach is historically the very first. This approach is based on the idea that we assume some distribution function for the data in the training set. This function has a set of theta parameters and our task is to evaluate this vector of parameters from the training set. The second approach is a bit more complicated. It often turns out that one model is not enough and we must use a mixture of models - the sum of several distribution functions and each of these functions has its own set of parameters. We will get acquainted with this model in the next lecture. The listed models are called parametric because they use a vector of parameters to configure them. The third method is called nonparametric and its idea is to restore the likelihood function without using anything at all except sample for training. It is with this method that we will begin.

Slide 17

What is the easiest way to estimate the probability distribution for a quantity without using complex functions with a parameter? Use a histogram! For a discrete random variable, the histogram will be an estimate of the frequency of occurrence of a certain value of a random variable over the training set. Let me remind you that square brackets mean Averson’s notation. If the condition in square brackets is satisfied, then this function will return the value 1, otherwise 0. But what about continuous descriptions of random variables? We can also use the histogram principle, but now we will evaluate the probability of the value of x using the counter of the number of examples of values ​​of this random variable in a certain range of values. Doesn’t resemble anything? That's right, we already saw it!

Slide 18

If we replace the function under the sum expression with any symmetric and normal function, then we get exactly the method of k-nearest neighbors, which we studied before. But this method is used here with the difference that the kernel function must be symmetrical and normalized. Otherwise, we will not be able to obtain an estimate of the probability density function.

Slide 19

Prior to this, we considered only a one-dimensional case. What to do with multidimensional distributions? First, we have already mentioned the naive Bayes classifier. Secondly, we can use some tricks. The first is presented at the beginning of the slide. We also use one-dimensional kernels to estimate the distribution density, but under the sign of the sum we multiply all such kernels for each feature. This is very similar to the naive Bayesian classifier, but not quite. The second method consists in using the same idea as in the method of nearest neighbors. Let's introduce a metric that allows us to estimate the distance between the objects of the training sample and use this distance to estimate to restore the multidimensional probability density. True, it is necessary to normalize to the sphere of unit radius according to this metric. This is an inconvenient value and it is difficult to evaluate, but as we will see now, it can be eliminated.

Slide 20

To begin, let's write an expression of likelihood assessment for a particular class. As you can see, this is not a big deal. You can simply substitute this likelihood estimate into the Bayesian classifier as is. Please note that due to the fact that we use the argmax expression, we can ignore the division by V (h), since this value does not depend on the class of the object. Thus, we were able to combine two worlds - the metric approach to classification and the approach using probability theory.

Slide 21

Let me remind you of a few things regarding the setting of a classifier that uses a metric approach. First, we need to choose a metric. The generalized expression of the metric in Euclidean space is presented on the slide. This metric is called the Minkowski weighted metric. It uses an arbitrary degree p for the difference in the values ​​of the features. Features are also weighted among themselves, which allows the use of such a metric to determine the degree of importance of the features and their selection. In general, the parameters of such a metric can be adjusted using the leave-one-out method and estimation of weights using linear models.

Slide 22

You can also choose different forms of kernels to estimate the probability density. Let me remind you that in general, this kernel does not affect the quality of classification in any way, and you can choose an arbitrary kernel. The only difference is that the density estimate will be more or less smooth.

Slide 23

So, we examined a nonparametric approach to estimating the probability distribution density. As I already said, there is also a parametric approach, which consists in using some kind of predetermined model for representing data and their distribution density and estimating the parameters of this model to restore the probability density function. This group of models includes both simple models and mixtures of models. In this lecture we will consider only simple models.

Slide 24

Before proceeding to consider a specific model, let's look at the most important principle, which will allow you to restore the parameters of any probabilistic model from the training sample. This principle is called the "maximum likelihood principle". It is based on a simple idea. Let's take some probabilistic model and a data sample that can be described by this model. Moreover, this sample was formed independently. Then the probability of obtaining such a sample can be estimated by a simple product of the probabilities of the individual elements. In our case, this probability is estimated by the model, so that the probability of observed sample can be estimated using the results of using the model. But this sample has already been observed, which means that most likely this is a more probable event and our model should return the maximum possible probability estimation for such sample. Thus, we can link the estimation of model parameters to the problem of maximizing the probability of the observed sample. This is what is called the maximum likelihood principle. According to this principle, the model parameters should be selected so as to maximize the probability estimate of the observed sample.

Slide 25

If we apply a logarithm function for the estimation of the probability of the observed sample, then we can obtain the sum of the logarithms of the individual values ​​of the objective function. One of the methods for obtaining the optimal values ​​of the model parameters, in which the maximum likelihood function is achieved, is to use the gradient descent method. To do this, you need to change the sign in the expression for evaluating the likelihood function and the maximization problem becomes a minimization problem. This is exactly how the model of logistic regression considered by us earlier works. The sigmoid that this model uses is a probability distribution model. If you carefully write the likelihood function using a sigmoid, you can get those expressions that we saw in the fourth lecture.

Slide 26

Let's look at another model of multidimensional probability density. This model is based on the Gaussian distribution function you know for a multidimensional random variable. In this model, it is assumed that all features are subsets of real numbers. If so, then you can use the expression on the slide. Sigma and mu in this expression are generalizations of the one-dimensional problem to the multidimensional case. The sigma covariance matrix contains estimates of the correlations of pairs of features among themselves. Det of matrix sigma is a matrix determinant. All these values ​​can be calculated using a training sample of a sufficiently large size and obtain optimal parameters for this model from the point of view of the maximum likelihood principle.

Slide 27

Let me remind you that the normal distribution in the multidimensional case can be represented as a cloud of points, which is compactly placed in this space and has a center of mass, in which the density of points is maximum. An example of points that are described by a multidimensional normal distribution is presented on the slide.

Slide 28

The model based on the normal distribution is very easy to consider and easy to understand, but it has a number of significant drawbacks. The first drawback is that the data in the training set can be represented so that the sigma matrix is ​​degenerate. This means that it will not be possible to calculate the inverse matrix of it, and its determinant is zero. This problem can be solved using regularization of the matrix and by getting rid of linearly dependent features. The second problem of estimating the normal distribution is that the parameters of this distribution are very sensitive to outliers. See how much the average estimate changes after adding just one outlier. This problem can be solved using more stable estimations, such as the median. Finally, the third and the most important problem is that the normal distribution is both a very simple model and a very specific one. If these tasks are far from the normal distribution, then such a model, in principle, cannot be used for correct estimates. A mixture of normal distributions, which we will discuss in the next lecture, helps to correct this situation. Now let's talk about how to regularize the covariance matrix.

Slide 29

One of the ideas of regularization is to observe that the covariance matrix becomes degenerate if for some particular measurements the point cloud has a very small spread of values. In other words, this cloud is very narrow. You can see such an example on the slide. If this cloud has a narrow shape, then this means that some eigenvalues ​​of the covariance matrix are close to zero. To fix this, you can explicitly increase them by some constant. For us, this can be represented as if a point cloud began to swell in all directions. To do this, it is enough to add this constant to all diagonal values ​​of the covariance matrix. Thus, the covariance matrix will be more like a diagonal matrix.

Slide 30

Another procedure can lead to reducing the matrix to diagonal form. It is possible to explicitly equate individual off-diagonal elements of this matrix to zero. To do this, we need to test the statistical hypothesis that such off-diagonal elements are equal to zero. For this, you can use the Student’s distribution. This distribution allows us to evaluate whether two random variables really have a correlation or whether this is a random coincidence. On the slide, you can see formulas to test this statistical hypothesis.

Slide 31

As I already said, for a method based on a multidimensional normal distribution, it is important to somehow get rid of outliers when estimating parameters. Another way to deal with them is to explicitly use the likelihood function. To do this, we must train the model in the sample and calculate the probabilities of each element of such a sample. Further, all the elements must be sorted in descending order of the likelihood function, and the right-hand side of this series will just contain outliers. Once outliers have been determined, they must be removed from the training sample and the model could be re-trained.

Slide 32

So, today we began to consider one of the most powerful approaches in machine learning based on the use of probability theory. There is a theorem that tells us that the Bayesian classifier is an optimal classifier. We learned what this classifier consists of and we examined two approaches to assessing its components - the non-parametric Parzen-Rosenblatt method and the parametric method using the multidimensional normal distribution. But what if we want to use the parametric method, but the multidimensional normal distribution is not suitable for us, because it poorly describes the data? In this case, using a mixture of normal distributions will help us. We will talk about this algorithm in the next lecture.