Вопросы на ”до свидания”

1. Полнота, точность

   Полнота, она же Recall - сколько фрода мы поймали.

   Точность, Precision - сколько не-фрода мы пропустили.

   $$Recall = \frac{tp}{tp + fn}$$

   $$Precision = \frac{tp}{tp + fp}$$

2. ROC кривая. Коэфициент Джини. AUC

   ROC кривая - зависимость true positive от false positive для разных порогов решающего правила.

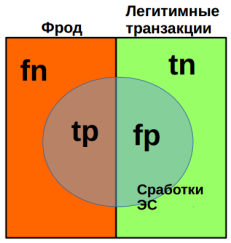
   Коэффициент Джини - максимальное отклонение ROC кривой от диагонали $tp = fp$.

   Если система адекватная (не контрпродуктивная), то кривая всегда будет выше диагонали.

   AUC - ”area under curve” - площадь под кривой ROC. Чем выше, тем лучше.

3. tp, fp, tn, fn

...	Фрод	Не фрод
Думаем, что фрод	tp	fp
Думаем, что не фрод	fn	tn



4. классификация

   Есть несколько категорий (напр. фрод\не фрод). Необходимо объекты расфасовать по категориям. В ”обучающей выборке” заранее известно, в какую категорию какой объект входит.

5. кластеризация

   Категории заранее неизвестны. Необходимо разбить объекты на несколько ”каких-то” категорий.

6. регрессия

   Зная значение функции на каком-то наборе точек, наиболее хорошо приблизить функцию.

7. идентификация

   Определить, к какой из большого числа категорий принадлежит объект.

8. обучение с учителем и обучение без учителя

   Supervised\Unsupervised learning. При обучении с учителем системе дается какая-то обучающая выборка, для которой задача уже решена (например, при задаче классификации). Без учителя обучающей выборки нет, система сама должна искать закономерности (кластеризация).

9. признак

   Строго заданная характеристика объекта (например, сумма транзакции в рублях).

10. класс

    Одна из категорий, принадлежность к которой необходимо определить при решении задачи классификации.

11. событие, действие, решение

    ???

12. обратная связь

    В общем случае: влияние выхода системы на ее дальнейший вход.

13. DENY, REVIEW, ALLOW решения ФМ системы RSA

    DENY - заблокировать транзакцию и профиль клиента.

    REVIEW - потребовать подтверждение в КЦ.

    ALLOW - разрешить без вопросов.

14. примеры применения машинного обучения для задач инфо

    Имея транзакции банка, определить фрод.

    Предсказать нагрузки на сайт в разное время, дабы не допустить DoS.

15. экспертная система

    ???

16. выброс

    Нестандартные данные, которые могут помешать классификации.

17. аномалия

    ???

18. выборка

    Конечный набор объектов, взятый из множества всех возможных объектов.

19. разделяющая гиперповерхность

    Гиперплоскость в пространстве всех возможных объектов, которая разделяет объекты на классы.

20. функция штрафа

    Некая функция, символизирующая, насколько "плохо" был классифицирован объект. Обычно функция расстояния или просто $+1$ за каждую точку не в своем классе. Задача классификации сводится к задаче минимизации функции штрафа.

21. логистическая функция

    $$f(\pm d(x_1, x_2)) = f(d) = \frac{1}{1 - e^{-\alpha \times d}}$$

    Используется в качестве функции штрафа.

22. ансамбль

    Мнения нескольких систем так или иначе аггрегируются в одно (например, функцией голосования).

23. дерево решений

    Дерево, в каждом узле которого находится условие, определяющее, к какому из детей надо перейти. В листьях содержатся классы. Каждый объект спускается по дереву в соответствии с условиями, пока не попадет в класс.

24. бутстрепинг

    Генерация выборки размера N из подвыборки размера $n \ll N$ путем выбора с повторением.

25. бутстреп аггрегация

    Она же bagging. Тот же ансамбль на классификаторах.

    Обобщение одного и того же классификатора на различных подвыборках или случайных параметрах, которые различны при каждой новой подгонки.

26. подгонка (fitting)

    Настройка параметров классификатора с помощью обучающих выборок.

27. проверка (scoring)

    Оценка качества классификатора на тестовой выборке.

28. переобучение (overfitting)

    Явление, когда модель слишком точно подогнана под обучающие данные, видя закономерности в случайном шуме. Такая модель может работать хуже на больших данных.

29. пакеты pandas, numpy, scikit-learn, matplotlib

    Библиотеки для Python.

    Pandas - структуры данных.

    numpy - матрицы, математика.

    scikit-learn - машинное обучение

    matplotlib - графики