Задание 4. Потапов Анатолий.

Теоретические задачи 3.1 Знакомство с линейным классификатором

1. Как выглядит бинарный линйеный классификатор?

   Положим $Y = \{-1, +1\}$ - можество классов, X - множество объектов, $x_1, ..., x_n$ - признаки объектов. Линейным классификатором называется алгоритм классификации $f(x) : X \to Y$ вида
   $$f(x, w) = sign\left(w^\top x + w_0\right)$$
   гдеw_0 - порог принятия решения, $w = (w_1, ..., w_n)^\top$ - вектор весов. $x_0 = 1$ полагается по умолчанию.

2. Что такое отступ алгоритма на объекте? Какие выводы можно сделать из знака отступа?

   В случае двухклассовой классификации отступом для объекта выборки $x_i \in X^n$ называетсяя величина
   $$M(x_i) = y_i f(x_i)$$
   где y_i - метка класса. Как видно из определения, отступ положителен, если алгоритм отнес объект к правильному классу (знаки y_i и $f(x_i)$ совпадают) и отрицателен, если алгоритм ошибся. Величина отступа интерпретируется как степень уверенности алготирма: чем больше отступ, тем более уверенная классификация.

3. Как классификаторы вида $a(x) = sign(\langle w, x \rangle - w_0)$ сводят к классификаторам вида $a(x) = sign(\langle w, x \rangle)$?

   Ответ: полагают $x_0 = -1$.

4. Как выглядит запись фукнционала эмпирического риска через отступы? Какое значение он должен принимать для "наилучшего"алгоритма классификации?

   Функция эмпирического риска есть
   $$Q(w) = \sum_{i=1}^{n} I\{M_i(w) < 0\}, \quad \text{где } M_i(w) \text{ - отступ на i-м объекте}$$
   I - индикатор.

   Для "найлучшего"алгоритма классификации функция импирического риска равна 0.

5. Если в функционале эмпирического риска (риск с пороговой функцией потерь) всюду написаны строгие неравенства ($M_i < 0$), можете ли вы сразу придумать параметр w для алгоритма классификации $a(x) = sign(\langle w, x \rangle)$, минимизирующий такой функционал?

   Ответ: $w = 0$?

6. Запишите функционал аннроксимирующего импирического риска, если выбрана функция потерь $L(M)$.
   $$Q^*(w) = \sum_{i=1}^{n} L(M_i(w)), \quad \text{где } M_i(w) \text{ - отступ на i-м объекте}$$

7. Что такое функция потерь и зачем она нужна? как обычно выглядит ее график?

   Функция потерь - это напрерывная или гладкая функция $L : \mathbb{R} \to \mathbb{R}_+$, что

   $$I\{M < 0\} \leqslant L(M)$$

   Функция потерь - это непрерывная аппроксимация пороговой функции. Как правило, это монотонная невозрастающая функция.

8. Приведите пример негладкой функции потерь

   $$L(x, y) = |x - y|$$

9. Что такое регуляризация? Какие регуляризаторы вы знаете?

   Регуляризация - это штраф за слощность модели.

   $$Q^*(w) = \sum_{i=1}^{n} L(M_i(w)) + \gamma R(w)$$

   $$R(w) = ||w||_1 - L_1 \text{ регуляризация}$$

   $$R(w) = ||w||_2 - L_2 \text{ регуляризация}$$

10. Как связаны переобучение и обобщающая способность алгоритма? Как влияет регуляризация на обобщающую способность?

    Переобучение - отсутствие обощающей способности алгоритма (сильная подстройка под данные, отсутствие выявления закономерностей). Регуляризация нужна для упрощения модели, для уменьшения подстройки под тренировочные данные и, как следствие, для увеличения обобщающей способности.

11. Для какого алгоритма классификации функционал риска будет принимать большее значение на обучающей выборке: для построенного с регуляризацией или без нее? Почему?

    Без регуляризации функционал риска будет меньше, чем с регуляризацией, так как будет бОльшая подстройка под тренировочные данные.

12. Для какого алгоритма классификации функционал риска будет принимать большее значение на тестовой выборке: для построенного с регуляризацией или без нее? Почему?

    Функционал риска на тестовой выборкебудет принимать меньшее значение, если он построен с регуляризацией. Без регуляризации алгоритм очень хорошо описывает объекты и обучающей выборки, но совсем не имеет обобщающей способности.

13. Что представляют собой метрики качества Accuracy, Precision и Recall?

    precision $= \frac{TP}{TP+FP}$, recall $= \frac{TP}{TP+FN}$, accuracy $= \frac{\sum TP + \sum TN}{\text{total number of objects}}$

    где TP - число объектов, класса 1, и классифицированных, как класс 1.

    FP - число объектов класса 0, классифицированных, как класс 1.

    FN - число объектов класса 1, отнесенных к классу 0.

    TN - число объектов класса 0, классифицированные, как класс 0.

14. Что такое метрика качества AUC и ROC-кривая?

    ROC - кривая - это кривая, показывающая зависимость доли верных положительных классификаций от доли ложных положительных классификаций при варьировании порога решающего правила.

    Как уже сказано, ROC-кривая строится в следующих координатах: по оси X - доля положительных ошибочных классификаций (false positive rate). Если N - число объектов класса 0, то $FPR = \frac{FP}{N} = \frac{FP}{FP+TN}$. По оси y - доля правильных положительных классификаций (true positive rate). $TPR = \frac{TP}{TP+FN}$. ROC-кривая выходит из точки (0;0) и заканчивается в точке (1;1).

    AUC - area under ROC curve - площадь под ROC кривой. Чем выше AUC, тем качественнее классификатор. Если AUC = 0.5, то классификатор просто пытается угадать ответы.

15. Как построить ROC-кривую, если у вас есть правильные ответы к домашнему заданию про фамилии?

    После обучения просим алгоритм выдать вероятности принадлежности слова к классу 1 (объекты класса 1 - фамилии, класса 0 -не фамилии). У каждого объекта есть предсказананя вероятность - на всей тестовой выборке получаем вектор (числа в нем из отрезка [0;1]). Дальше, варьируя величину порога t решающего правила (например если значение предсказанной вероятности больше $t = 0,6$, то относим объект к классу 1, а если $< t$, то к классу 0), считаем метрики FPR и TPR (см. предыдущий пункт) и откладываем их по осям. Порог пробегает значение из отрезка [0,1].

3.2 Вероятностный смысл регуляризаторов

Покажите, что регуляризатор в задаче линейной классификации имеет вероятностный смысл априорного распределения параметров моделей. Какие распределения задают l_1- и l_2-регуляризаторы.

Запишем постановку задачи классификации:

$$Q = \sum_{i=1}^{\ell} L(y_i, f(x_i)) + \gamma V(w) \to \min_w$$

где L - заданная функция потерь, $V(w)$ - регуляризатор.

Исходную задачу можно переписать так:

$$\sum_{i=1}^{\ell} -L(y_i, f(x_i)) - \gamma V(w) \to \max_w$$

$$\sum_{i=1}^{\ell} \ln e^{-L(y_i, f(x_i))} + \ln e^{-\gamma V(w)} \to \max_w$$

Воспользуемся свойствами логарифма и перепишем задачу так:

$$e^{-\gamma V(w)} \prod_{i=1}^{\ell} \ln e^{-L(y_i, f(x_i))} \to \max_w$$

Функция $e^{-\gamma V(w)}$ задает априорое распределение параметров модели, а именно коэффициентов весов.

Если $V(w) = ||x||_1$, то $P(w)\ e^{-\gamma ||w||_1}$ - похоже на плотность лапласовского распределения $p(x) \sim e^{-||x||_1}$).

Если $V(w) = ||x||_2$, то $P(w)\ e^{-\gamma ||w||_2}$ - похоже на плотность гауссовского распределения $(p(x) \sim e^{-x^\top x})$.

3.3 SVM и максимизация разделяющей полосы Покажите, как получается условная оптимизационная задача, решаемая в SVM из соображений максимизации разделяющей полосы между классами. Можно отталкиваться от линейно разделимого случая, но итоговое решение должно быть для общего.

Как эта задача сводится к безусловной задаче оптимизации?

Решающая функция имеет вид

$$f(x) = sign\left(\langle w, x \rangle - w_0\right)$$

Ясно, что $\langle w, x \rangle - w_0 = 0$ - это уравнение разделяющей гиперплоскости.

Умножив, если необходимо, функцию $f(x)$ на положительную константу, считае далее, что $\min\limits_{i=1,\dots,\ell} y_i \left(\langle w, x_i \rangle - w_0\right) = 1$. Тогда ясно, что $y_i \cdot \left(\langle w, x_i \rangle - w_0\right) \geqslant 1$ - ни одна точка не может лежать внутри разделяющей полосы (при линейно разделимой выобрке). Отсюда следует, что $-1 < \langle w, x_i \rangle - w_0 < 1$ - полоса, разделяющая классы.

Обозначим объекты, лежащие на границе разделяющей полосы максимальной ширины x_- и x_+. Уравнения разделяющих гиперплоскостей, проходящих через эти точки, есть $\langle w, x_+ \rangle - w_0 = 1$ и $\langle w, x_+ \rangle - w_0 = -1$ Тогда расстояние между этими крайними положениями гиперплоскости есть

$$\left\langle (x_+ - x_-), \frac{w}{||w||} \right\rangle = \frac{\langle x_+, w \rangle - \langle x_-, w \rangle}{||w||} = \frac{(w_0 + 1) - (w_0 - 1)}{||w||} = \frac{2}{||w||}$$

Таким образом, чтобы максимизировать ширину разделяющей полосы, надо минимизировать $||w||$.

Получили следующую задачу условной оптимизации:

$$w^\top w \to \min$$

$$\text{s.t.} y_i \cdot \left(\langle w, x_i \rangle - w_0\right) \geqslant 1$$

В случае линейно неразделимой выборки можно ввести неотрицательные штрафы ξ_i за попадание объекта из тестовой выборки в разделяющую полосу. Естественно, за попадание в разделяющую полосу нужно штрафовать. Поэтому задача условной оптимизации для линейно неразделимой выборки имеет следующий вид:

$$w^\top w + C \sum_{i=1}^{\ell} \xi_i \to \min_{w, w_0, \xi}$$

$$\text{s.t.} y_i \cdot (\langle w, x_i \rangle - w_0) \geqslant 1 - \xi_1, \quad i = 1, ..., \ell$$
$$\xi_i \geqslant 0, \quad i = 1, ..., l.$$

Покажем, как свести полученную задачу к задаче безусловной оптимизации. Вспомним, что отступом классификатора на объекте y_i называется величина $M_i = y_i(\langle w, x_i \rangle - w_0)$. Тогда два условия из задачи условной оптимизации можно записать как одно условие, используя понятие отступа:

$$\begin{cases} \xi_i \geqslant 0 \\ M_i \geqslant 1 - \xi_i \end{cases} \Rightarrow \begin{cases} \xi_i \geqslant 0 \\ \xi_i \geqslant 1 - M_i \end{cases} \Rightarrow \xi_i = \max\{0, 1 - M_i\} = (1 - M_i)_+$$

(оператор $(\cdot)_+$ называют положительной срезкой.

Теперь задача имеет виду безусловной задачи оптимизации:

$$Q(w, w_0) = \sum_{i=1}^{\ell} (1 - M_i(w, w_0))_+ + \frac{1}{2C} ||w||^2 \to \min_{w, w_0}$$

3.4 Kernel trick Придумайте ядро, которое позволит линейному классификатору с помощью Kernel Trick построить в исходном пространстве признаков разделяющую поверхность $x_1^2 + 2x_2^2 = 3$. Какой будет размерность спрямляющего пространства?

Найдем $\phi : \mathbb{R}^2 \to \mathbb{R}^3$. Положим

$$\phi(x_1, x_2) = (x_1, x_2, x_1^2 + 2x_2^2 - 3)^\top$$

График ϕ - параболоид. Уравнение $\phi(x_1, x_2) = 0$ в точности задает уравнение элиппса: $x_1^2 + 2x_2^2 = 3$, что и требовалось. Спрямляющее пространство - плоскость, ее размерность - 2.

3.5 ℓ_1-регуляризация

Покажите с помощью теоремы Куна-Таккера, что ограничение ℓ_1-нормы вектора весов числом и добавление штрафа с его ℓ_1-нормой приводят к построению одного и того же алгоритма. Можно считать, что регуляризатор добавляется по существу, т.е. меняет итоговый ответ по сравнению с оптимизационной задачей без регуляризатора.

Покажем это на примере классификатора (без ограницения общности, это можно показать и для других алгоритмов). Запишем задача классификации с ограничением ℓ_1-нормы вектора весов числом в каноническом виде:

$$\sum_{i=1}^{\ell} Loss(M_i(w)) \to \min_w$$

$$\text{s.t.} ||w||_1 - \mathrm{C} \leqslant 0$$

Теорема Каруша-Куна-Таккера дает необходимые (а в случае выпуклой задачи и достаточные) условия оптимальности:

$$\nabla_w L(w, \mu) = \nabla_w \left(\sum_{i=1}^{\ell} Loss(M_i(w)) + \mu(||w||_1 - C) \right) = 0$$

$$\mu \cdot (||w||_1 - C) = 0$$

$$\mu \geqslant 0$$

Перепишем первое условие:

$$\nabla_w L(w, \mu) = \nabla_w \left(\sum_{i=1}^{\ell} Loss(M_i(w)) + \mu ||w||_1 - \mu C \right) = 0$$

Получили, что к минимизируемой функции прибавился штраф в виде ℓ_1 нормы вектора весов. Значит (если функция потерь $Loss$ выпуклая, то эти две задачи эквивалентны, ч.т.д.

3.6 Повторение: метрики качества. Уже сделано, см. номер **3.1**