Знакомство с линейным классификатором

1. Как выглядит бинарный линейный классификатор? (Формула для отображения из множества объектов в множество классов.)

Бинарный линейный классификатор $a : X \to \{-1, +1\}$ действует по формуле

$$a(x) = \text{sign}\,(f(x))\,,$$

где $f(x) = w_0 + \langle w, x\rangle$ – некоторая линейная функция.

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

Отступом алгоритма $a(x) = \text{sign}(f(x))$ на объекте x_i называется величина $M_i = y_i f(x_i)$, где $y_i = \pm 1$ – класс объекта x_i. Если M_i положительно, то $y_i = a(x_i)$, то есть классификатор выдаёт правильный ответ. А если M_i отрицательно, то $y_i \neq a(x_i)$, и классификатор ошибается.

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

Добавим объектам новый признак, тождественно равный -1.

4. Как выглядит запись функционала эмпирического риска через отступы? Какое значение он должен принимать для «наилучшего» алгоритма классификации?

Функционал эмпирического риска: $Q(w) = \sum\limits_i I(M_i(w) < 0)$. Для «наилучшего» алгоритма классификации он должен принимать значение 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 $a(x) = \text{sign}(\langle w, x\rangle)$, минимизирующий такой функционал?

Да. Если $w = 0$, то и $f \equiv 0$, следовательно $M_i = 0$ для любого i, а потому $Q(w) = 0$.

6. Запишите функционал аппроксимированного эмпирического риска, если выбрана функция потерь $L(M)$.

$$\tilde{Q}(w) = \sum_i L(M_i(w))$$

7. Что такое функция потерь, зачем она нужна? Как обычно выглядит ее график?

Минимизация функционала эмпирического риска по вектору весов сводится к поиску максимальной совместной подсистемы в системе неравенств. А это NP-трудная задача. Однако для практического интереса бывает достаточно приближённого решения, достаточно близкого к точному. Для этого мы заменяем пороговую функцию потерь $I(M < 0)$ её аппроксимацией $L(M)$, где $M : \mathbb{R} \to \mathbb{R}_+$. Причём, M является непрерывной, как правило, гладкой и $I(M < 0) \leqslant M(L)$. Тогда $Q(w) \leqslant \tilde{Q}(w)$. Вместо минимизации функционала Q происходит минимизация функционала $\tilde{Q}$.

8. Приведите пример негладкой функции потерь.

$$\max(1 - M, 0)$$

9. Что такое регуляризация? Какие регуляризаторы вы знаете?

Регуляризация – это добавление к минимизируемому функционалу некоторого штрафного слагаемого, запрещающее слишком большие значения весов. Например, l_p-регуляризация – минимизация функционала

$$\tilde{Q}(w) = \sum_i L(M_i(w)) + \gamma \sum_k w_k^p$$

Регуляризация снижает риск переобучения и повышает устойчивость вектора весов по отношению к малым изменениям обучающей выборки.

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

Чем выше обобщающая способность, тем меньше риск переобучения. Регуляризация улучшает обобщающую способность.

11. Как связаны острые минимумы функционала аппроксимированного эмпирического риска с проблемой переобучения?

Острые минимумы функционала неустойчивы. Поэтому попадание в один из таких минимумов может вести к переобучению.

12. Что делает регуляризация с аппроксимированным риском как функцией параметров алгоритма?

Регуляризация повышает устойчивость решения w, тем самым улучшая обобщающую способность алгоритма.

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

Конечно, без регуляризации. Потому что в одном случае происходит просто минимизация функционала на обучающей выборке, а в другом случае минимизация происходит под ограничением регуляризатора.

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

С регуляризацией. Иначе она не была бы «оправдывающей себя». Если алгоритм на тестовой выборке без регуляризации работает лучше, то спрашивается: а зачем нам такая регуляризация сдалась?

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

Бинарный классификатор может выдать 2 типа ответов: положительный и отрицательный. Разобьём решения классификатора на 4 типа: TP – истинно-положительные решения, TN – истинно-отрицательные решения, FP — ложно-положительные решения, FN — ложно-отрицательные решения. Тогда

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \text{ – доля правильных ответов;}$$

$$\text{Precision} = \frac{TP}{TP + FP};$$

$$\text{Recall} = \frac{TP}{TP + FN}.$$

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

Пусть классификатор представлен в виде $a(x) = \text{sign}(f(x, w) - w_0)$, где x – объект, $f(x, w)$ — некоторая функция, w — вектор параметров, определяемый по обучающей выборке, w_0 — порог. Определим две характеристики

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

ROC-кривой называется параметрическая кривая $(FPR(w_0), TPR(w_0))$. Она начинается в точке $(0, 0)$, заканчивается в точке $(1, 1)$ и задаёт график монотонно неубывающей функции. Метрика качества AUC есть просто площадь под ROC-кривой.

17. Как построить ROC-кривую (нужен алгоритм), если например, у вас есть правильные ответы к домашнему заданию про фамилии и ваши прогнозы?

1) Вычислим количество представителей классов в выборке, то есть величины $TP + FP$ и $TN + FN$. (Они зависят только от выборки, а не от порога w_0.)

2) Упорядочим выборку по убыванию значений $f(x_i, w)$.

3) Начальную точку устанавливаем $(FPR_0, TPR_0) = (0, 0)$.

4) Следующие точки вычисляем рекуррентно. Если объект x_i относится к положительному классу, то $FPR_i = FPR_{i-1}$, $TPR_i = TPR_{i-1} + \frac{1}{TP+FP}$. Иначе $FPR_i = FPR_{i-1} + \frac{1}{TN+FN}$, $TPR_i = TPR_{i-1}$.

5) Последнюю точку устанавливаем $(FPR_{m+1}, TPR_{m+1}) = (1, 1)$.

Вероятностный смысл регуляризаторов

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

Допустим, что множество $X \times Y$ является вероятностным пространством, причём распределение объектов и классов задано совместной плотностью $p(x, y \mid w)$, зависящей от вектора параметров w. Более того, допустим, что также имеется априорное распределение в пространстве параметров модели $p(w)$. Тогда логарифмическая функция правдоподобия равна

$$\sum_i \ln p(x_i, y_i \mid w) + \ln p(w).$$

Тогда понятно, что если положить

$$-\ln p(x_i, y_i \mid w) = L(y_i f(x_i, w)),$$

то по принцип максимального правдоподобия приходим к задаче

$$\sum_i L(y_i f(x_i, w)) - \ln p(w) \to min,$$

то есть минимизация функционала аппроксимированного эмпирического риска с регуляризатором $-\ln p(w)$.

Если вектор w имеет нормальное распределение, все его компоненты независимы и имеют равные дисперсии σ, то

$$\ln p(w,\sigma) = \ln\left(\frac{1}{(2\pi\sigma)^{\frac{n}{2}}}\exp\left(-\frac{||w||_2^2}{2\sigma}\right)\right) = -\frac{1}{2\sigma}||w||_2^2 + const$$

и получается l_2-регуляризация.

Если вектор w имеет априорное распределение Лапласа, все его компоненты независимы и имеют равные дисперсии, то

$$\ln p(w,C) = \ln\left(\frac{1}{(2C)^n}\exp\left(-\frac{||w||_1}{C}\right)\right) = -\frac{1}{C}||w||_1 + const$$

и получается l_1-регуляризация.

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 выражение должно быть для общего. Как эта задача сводится к безусловной задаче оптимизации?

Итак, мы строим линейный пороговый классификатор:

$$a(x) = \text{sign}\left(\langle w, x\rangle - w_0\right)$$

Сначала предположим, что выборка линейно разделима. Тогда функционал числа ошибок принимает нулевое значение, и разделяющая гиперплоскость не единственна. Заметим, что параметры линейного классификатора определены с точностью до нормировки. Умножим w и w_0 на одну и ту же положительную константу так, чтобы выполнялось условие

$$\min_i y_i\left(\langle w, x\rangle - w_0\right) = 1.$$

Множество точек $\{x : |\langle w, x\rangle - w_0| \leqslant 1\}$ описывает полосу, разделяющую классы. Ни один из объектов обучающей выборки не попадает внутрь этой полосы. Границами полосы служат две параллельные гиперплоскости с вектором нормали w. Разделяющая гиперплоскость проходит ровно по середине между ними. Идея состоит в том, чтобы максимизировать ширину этой полосы. А ширина такой полосы равна $\frac{2}{||w||_2}$. В итоге, в случае линейно разделимой выборки получаем задачу условной оптимизации:

$$\begin{cases}\langle w, w\rangle \to \min;\\ y_i\left(\langle w, x\rangle - w_0\right) \leqslant 1.\end{cases}$$

Чтобы обобщить постановку задачи на случай линейно неразделимой выборки, позволим алгоритму допускать ошибки на обучающих объектах, но при этом постараемся, чтобы ошибок было поменьше. Введём дополнительные переменные $\xi_i \geqslant 0$, характеризующие величину ошибки на объектах x_i.

Ослабим в ограничения-неравенства и одновременно введём в минимизируемый функционал штраф за суммарную ошибку. Получается условная оптимизационная задача:

$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum\limits_i \xi_i \to \min\limits_{w, w_0, \xi}; \\ y_i\left(\langle w, x\rangle - w_0\right) \leqslant 1 - \xi_i; \\ \xi_i \geqslant 0. \end{cases}$$

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?

Итак, $X = \mathbb{R}^2$. Рассмотрим ядро $K(x, y) = \langle x, y\rangle^2$. Пусть $x = (x_1, x_2)$, $y = (y_1, y_2)$. Тогда

$$K(x,y) = (x_1y_1 + x_2y_2)^2 = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1y_1x_2y_2 = \left\langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (y_1^2, y_2^2, \sqrt{2}y_1y_2)\right\rangle$$

Ядро K представляется в виде скалярного произведения в пространстве $\mathbb{R}^3$. Преобразование $\psi : \mathbb{R}^2 \to \mathbb{R}^3$ имеет вид $\psi : (x_1, x_2) \to (x_1^2, x_2^2, \sqrt{2}x_1x_2)$. Линейной поверхности $z_1 + 2z_2 = 3$ в пространстве $\mathbb{R}^3$ соответствует квадратичная поверхность $x_1^2 + 2x_2^2 = 3$ в исходном пространстве. Таким образом, размерность спрямляющего пространства получилась равна 3.

Повторение: метрики качества

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

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

3. Как построить ROC-кривую (нужен алгоритм), если например, у вас есть правильные ответы к домашнему заданию про фамилии и ваши прогнозы?

Ответы на эти вопросы уже были даны ранее.