3.1 Знакомство с линейным классификатором

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

- Ответ: $\hat{y} = sign(+ w_0)$

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

- Ответ: отступ алгоритма $a(x) = sign(f(x))$ на объекте x_i это $M_i = y_i f(x_i)$

3. Как классификаторы вида a(x) = sign(< w, x > −w0) сводят к классификаторам вида a(x) = sign(< w, x >)?

- Ответ: достаточно добавить в x фиктивный константный столбец

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

- Ответ: $Q(w) = \Sigma\,[M_i <= 0]$, где $L(t)$ - функция потерь. Идеально 0.

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

- Ответ: w = 0

6. Запишите функционал аппроксимированного эмпирического риска, если выбрана функ- ция потерь L(M).

- Ответ: $\hat{Q}(w) = \Sigma L(M_i)$

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

- Ответ: функция потерь - штраф за неправильный ответ на объекте, нужна для оптимизации алгоритма, стараются минимизировать среднее её значение на обучающей выборке.

8. Приведите пример негладкой функции потерь.

- Ответ: $L(M) = (1 - M)_+$

9. Что такое регуляризация? Какие регуляризаторы вы знаете?

- Ответ: регуляризация - способ борьбы с переобучением, минимизируется не функция риска, а сумма функции риска и $\alpha R(w)$, где $R(w)$ - регуляризатор (например модуль или длинна вектора). l1, l2

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

- Ответ: чем больше переобучение тем меньше обобщающая способность и наоборот. Регуляризация увеличивает обобщающую способность

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

- Ответ: Острые минимумы функционала эмпирического риска говорят о том, что функция риска сильно меняется даже при небольшом изменении признаков. Из этого можно сделать вывод, что веса модели велики по значению и присутствует явление переобучения.

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

- Ответ: уменьшает переобучение (немного увеличивает риск)

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

- Ответ: с регуляризацией будет больше, в этом и весь смысл, чтобы алгоритм не переобучился

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

- Ответ: с регуляризацией меньше, в этом и весь её смысл. Потому что регуляризация не даст переобучиться на обучающей выборке

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

- Ответ: accuracy - доля правильных ответов, precision - правильные ответы в классе / размер предсказанного класса (fp/(tp + fp)), recall - правильные ответы в классе / размер класса (tp/(tp + fn))

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

- Ответ: AUC - площадь под ROC. ROC - кривая, которая строится так: для задачи бинарной классификации варьируется параметр w_0. Затем на график наносят точки (FPR, TPR), где TPR = TP / (TP + FN), FPR = FP / (FP + TN)

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

- Ответ: Пусть прогнозы -- вероятности. Тогда пробежимся порогом-отсечкой w_0 от нуля до единицы и каждый раз будем предсказывать ответы. Затем посчитаем TPR, FPR и нанесем их на график. Если прогнозы не вероятности достаточно посчитать для порогов 0, 1, и каким-нибудь одним между ними

3.2 Вероятностный смысл регуляризаторов

Покажите, что регуляризатор в задаче линейной классификации имеет вероятностный смысл априорного распределения параметров моделей. Какие распределения задают l1-регуляризатор и l2-регуляризатор?

Решение

Допустим, множество $X \times Y$ является вероятностным пространством, и задана параметрическая модель совместной плотности распределения объектов и классов $p(x, y|w)$. Введем параметрическое семейство априорных распределений $p(w; \gamma)$, где γ — неизвестная и не случайная величина (гиперпараметр).

Тогда будем считать, что выборка может быть порождена каждой из плотностей $p(x, y|w)$ с параметризованной γ вероятностью $p(w; \gamma)$.

Приходим к принципу максимума совместного правдоподобия данных и модели:

$$L_\gamma(w, X^l) = \ln p(X^l, w; \gamma) = \sum_{i=1}^{l} \ln p(x_i, y_i|w) + \underbrace{\ln p(w; \gamma)}_{\text{регуляризатор}} \to \max_w .$$

Вспомним, что этот принцип эквивалентен принципу минимизации аппроксимированного эмпирического риска

$$Q(w; X^l) = \sum_{i=1}^{l} \mathscr{L}(y_i f(x_i, w)) + \underbrace{\gamma V(w)}_{\text{регуляризатор}} \to \min_w,$$

если положить

$$-\ln p(x_i, y_i|w) = \mathscr{L}(y_i f(x_i, w)),$$
$$\ln p(w; \gamma) = \gamma V(w).$$

Таким образом, получаем, что регуляризатор $V(w)$ соответствует параметрическому семейству априорных распределений плотностей $p(w; \gamma)$ — параметров моделей.

ℓ_1-регуляризатор

Пусть $w \in \mathbb{R}^n$ имеет n-мерное распределение Лапласа:

$$p(w; C) = \frac{1}{(2C)^n} \exp\left(-\frac{\|w\|_1}{C}\right), \quad \|w\|_1 = \sum_{j=1}^{n} |w_j|,$$

т.е. все веса независимы, имеют нулевое матожидание и равные дисперсии; C — гиперпараметр.

Логарифмируя, получаем регуляризатор по ℓ_1-норме:

$$-\ln p(w; C) = \frac{1}{C} \sum_{j=1}^{n} |w_j| + \mathrm{const}(w).$$

ℓ_2-регуляризатор

Пусть $w \in \mathbb{R}^n$ имеет n-мерное гауссовское распределение:

$$p(w; \sigma) = \frac{1}{(2\pi\sigma)^{n/2}} \exp\left(-\frac{\|w\|^2}{2\sigma}\right), \quad \|w\|^2 = \sum_{j=1}^{n} w_j^2,$$

т.е. все веса независимы, имеют нулевое матожидание и равные дисперсии σ; σ — гиперпараметр.

Логарифмируя, получаем регуляризатор по ℓ_2-норме:

$$-\ln p(w; \sigma) = \frac{1}{2\sigma} \|w\|^2 + \mathrm{const}(w).$$

3.3 SVM и максимизация разделяющей полосы

Покажите, как получается условная оптимизационная задача, решаемая в SVM из соображе- ний 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?

Решение

Рассмотрим задачу классификации на два непересекающихся класса, в которой объекты описываются n-мерными вещественными векторами: $X = \mathbb{R}^n, Y = \{-1, +1\}$.

Будем строить линейный пороговый классификатор:

$$a(x) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = \operatorname{sign}\left(\langle w, x\rangle - w_0\right),$$

где $x = (x^1, \dots, x^n)$ — признаковое описание объекта x, вектор $w = (w^1, \dots, w^n) \in \mathbb{R}^n$ и скалярный порог $w_0 \in \mathbb{R}$ — параметры алгоритма.

Для начала предположим, что выборка линейна разделима: найдутся w, w_0, задающие разделяющую гиперплоскость $\langle w, x\rangle = w_0$, при которых функционал числа ошибок

$$Q(w, w_0) = \sum_{i=1}^{l} [y_i(\langle w, x_i\rangle - w_0) \leqslant 0] = 0.$$

Найдем, как оптимальнее расположить разделяющую гиперплоскость. Для простоты выполним нормировку параметров алгоритма: домножим w и w_0 на такую константу, что

$$\min_{i=\overline{1,l}} y_i \left(\langle w, x_i\rangle - w_0\right) = 1.$$

Хочется максимизировать ширину разделяющей полосы. Тогда на границе разделяющей полосы будут лежать точки из обучающей выборки: x_- и x_+, принадлежащие соответственно -1 и $+1$ классам. Ширина полосы

$$\left\langle (x_+ - x_-), \frac{w}{\|w\|} \right\rangle = \frac{\langle w, x_+\rangle - \langle w, x_-\rangle}{\|w\|} = \Big| y_i \left(\langle w, x_i\rangle - w_0\right) = 1, i \in \{+, -\} \Big| =$$
$$= \frac{(w_0 + 1) - (w_0 - 1)}{\|w\|} = \frac{2}{\|w\|}.$$

Получаем, что ширина полосы максимальна, когда норма вектора w минимальна. Значит, можно сформулировать следующую задачу оптимизации:

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min\limits_{w, w_0}; \\ y_i \left(\langle w, x_i\rangle - w_0\right) \geqslant 1, i = 1, \dots, l. \end{cases}$$

Если работать с линейно неразделимой выборкой, $y_i \left(\langle w, x_i\rangle - w_0\right)$ не обязательно будет не меньше 1. Ослабим эти ограничения и введем в минимизируемый функционал штраф за суммарную ошибку:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{l} \xi_i \to \min\limits_{w, w_0, \xi}; \\ M_i(w, w_0) = y_i \left(\langle w, x_i\rangle - w_0\right) \geqslant 1 - \xi_i, i = 1, \dots, l; \\ \xi_i \geqslant 0, i = 1, \dots, l. \end{cases}$$

ξ_i в этих условиях будет показывать величину ошибки на x_i объекте.

Преобразуем условия на ξ_i:

$$\begin{cases} \xi_i \geqslant 1 - M_i(w, w_0) \\ \xi_i \geqslant 0 \end{cases}$$

Значит, минимум ξ_i будет при $\xi_i = (1 - M_i(w, w_0))_+$.

Получаем эквивалентную задачу безусловной оптимизации:

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

3.4 Kernel trick

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

Решение

Возьмем квадратичное ядро: $K(x,y) = \langle x,y\rangle^2 = (x_1y_1 + x_2y_2)^2 = x_1^2y_1^2 + 2x_1y_1x_2y_2 + x_2^2y_2^2 = \langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (y_1^2, y_2^2, \sqrt{2}y_1y_2)\rangle$.

Получим отображение в спрямляющее пространство $H = \mathbb{R}^3$:

$$\psi : (x_1, x_2) \to (x_1^2, x_2^2, \sqrt{2}x_1x_2).$$

Тогда линейная поверхность в H будет иметь вид: $\langle (x_1^2, x_2^2, \sqrt{x}_1x_2), (w_1, w_2, w_3)\rangle + w_0 = w_1x_1^2 + w_2x_2^2 + w_3\sqrt{x}_1x_2 + w_0 = \Big| w = (1,2,0), w_0 = -3 \Big| = x_1^2 + 2x_2^2 - 3 = 0$.

(Вообще-то говоря, w ищется из условия $\dfrac{\partial \mathscr{L}}{\partial w} = w - \sum\limits_{i=1}^{l} \lambda_i y_i \psi(x_i) = 0$ после того, как найдены λ_i в задаче (4.1) со скалярным произведением $K(x_i, x_j)$ вместо $\langle x_i, x_j\rangle$.

w_0 можно найти, подставив в выражение $w_0 = \langle w, \psi(x_i)\rangle - y_i$ произвольный опорный граничный вектор x_i.)

3.5 l1-регуляризация

Покажите с помощью теоремы Куна-Таккера, что ограничение l1-нормы вектора весов числом и добавление штрафа с его l1-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.

Решение

Лассо Тибширани:

$$\begin{cases} Q(\alpha) = \|F\alpha - y\|^2 \to \min\limits_{\alpha}; \\ \sum\limits_{j=1}^{n} |\alpha_j| \leqslant \varkappa \;\Leftrightarrow\; \sum\limits_{j=1}^{n} |\alpha_j| - \varkappa \leqslant 0. \end{cases}$$

По теореме Куна-Таккера:

$$\begin{cases} Q(\alpha) = \|F\alpha - y\|^2 + \lambda\left(\sum\limits_{j=1}^{n} |a_j| - \varkappa\right) = \|F\alpha - y\|^2 + \lambda\left(\sum\limits_{j=1}^{n}\right)|a_j| + \mathrm{const} \to \min\limits_{\alpha}; \\ \lambda \geqslant 0; \\ \lambda\left(\sum\limits_{j=1}^{n} |\alpha_j| - \varkappa\right) = 0 \;\Leftrightarrow\; \lambda = 0 \text{ или } \sum\limits_{j=1}^{n} |\alpha_j| = \varkappa. \end{cases}$$

3.6 Повторение: метрики качества

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

- Ответ: accuracy - доля правильных ответов, precision - правильные ответы в классе / размер предсказанного класса (fp/(tp + fp)), recall - правильные ответы в классе / размер класса (tp/(tp + fn))

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

- Ответ: AUC - площадь под ROC. ROC - кривая, которая строится так: для задачи бинарной классификации варьируется параметр w_0. Затем на график наносят точки (FPR, TPR), где TPR = TP / (TP + FN), FPR = FP / (FP + TN)

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

- Ответ: Пусть прогнозы -- вероятности. Тогда пробежимся порогом-отсечкой w_0 от нуля до единицы и каждый раз будем предсказывать ответы. Затем посчитаем TPR, FPR и нанесем их на график. Если прогнозы не вероятности достаточно посчитать для порогов 0, 1, и каким-нибудь одним между ними

```
In [ ]:
```