Билет 1

Модель нейрона Маккалока-Питтса. Перцептрон Розенблатта. Теорема Новикова. Полносвязные нейронные сети

1) $X = (x_1, \dots x_n)$, нейрон $a(x)$ вычисляет n-арную булеву функцию: $a(x) = \text{Heaviside}(\sum_{j=1}^{n} w_j x^j - w_0)$, w - веса ($w_j > 0$ - возб., иначе - тормозящий)

2) Виды элементов: S - сенсорный эл-т, A - ассоциативный эл-т, ему соотв. несколько S - элементов. A активируется, если количество сигналов S на его входе превысило некоторое значение Θ. Затем сигналы ~~идут~~ от возбуждающих A идут на сумматор R с коэф. w_j - весом A-R связи (веса S-A принимают знач. -1, 0 или 1) значение порогов A рандомно и неизменно, общ. вид ф-ции, реализуемой R эл-том, представим как перцептрон $(x) = \text{sgn}(\sum_{j=0}^{n} w_j x_j - \Theta)$

3) Th Новикова и правило Хебба: Пусть мн-во прецедентов $Y = \{-1, +1\}$, x - объект обуч. выборки $X^\ell = \{x_j, y_j\}_{j=1}^{\ell}$, $y_i = y^*(x_j) \in Y$ - класса. Алг. класиф. имеет вид $a(x, w) = \text{sgn}(\langle x, w \rangle)$, ошибка возникает, если знак $\langle , \rangle$ не совпал со знаком класса, т.е. $\langle x, w \rangle y < 0$. Можно модифицировать веса: $w \to w + \eta x y$

Собственно теорема: $X = \mathbb{R}^{n+1}$, $Y = \{-1, 1\}$, выборка X^ℓ лин. разделима, т.е. $\exists w^*$ и $\delta > 0$ т.ч. $\langle x_j, w^* \rangle y_j > \delta \ \forall j = 1, \dots \ell$. Тогда алгоритм может найти вектор весов, разд. train без ошибок за кон. число итераций $\forall$ нач. приближ. w_0 и $\eta > 0$.

* FCN: input → x → $xW + b$ → σ → output

dense layer - лин. преобр. входн. данных (обуч. параметрами W и вектор b): $x \mapsto xW + b$, $W \in \mathbb{R}^{d \times k}$, $x \in \mathbb{R}^d$, $b \in \mathbb{R}^k$

Слой делает из d-мерных векторов k-мерные.

activation - нелин. преобр.

Билет 2

Аппрокс. теорема (без Δ): Лузина, Колмогорова-Арнольда, Цыбенко

1) Лузина: $\forall$ измеримой f на $[a, b]$ и $\forall \varepsilon > 0$ $\exists f_\varepsilon \in C[a, b]$, такая что $f = f_\varepsilon$ всюду кроме некоторого мн-ва меры $\leq \varepsilon$.

т.е. $M\{x \in [a,b] : f(x) \neq f_\varepsilon(x)\} < \varepsilon$

② Теорема Колмогорова - Арнольда: Всякая непр. ф-ция представима в виде композиции непр. ф-ций одной пер.

т.е. $\forall f(x)$, $x = (x_1, \dots x_n)$ $\exists$ представл. $\sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \Phi_{q,p}(x_p) \right)$

Уточнение Шпрехера: $f = \sum_{q=0}^{2n} \Phi\left(\sum_{p=1}^{n} \lambda_p \Phi(x_p + \gamma q) + q \right)$, $\gamma, \lambda \in \mathbb{R}$, $\Phi \in C$,

$\Phi \in R \to R$, $\Phi[0,2] \to R$.

③ Цыбенко: $\forall f \in C^0(R)$, $f : R \to R$, $\forall \varepsilon > 0 \; \exists N, w_1 \dots w_N, b_1 \dots b_N$,

$\alpha_1 \dots \alpha_N$: $|f - \sum_{i=1}^{} \alpha_i \sigma(\langle x, w_i \rangle + b_i)| < \varepsilon \; \forall x \in [0,1]^m \in \mathbb{R}^m$

Билет 3

Обучение нейросетей. Нейроном. stochastic gradient descent, back-propagation.

Минимизируем: $Q(w) = \sum h(w, x_i, y_i) \to \min$

$w = w - \eta \nabla h_i(w)$, $\nabla h_i(w) = \left(\frac{\delta h_i(w^k)}{\delta w^k} \right)_{k=1}^{m}$, $Q = (1-\lambda) Q + \lambda h_i(w)$

Backdrop

Дано $\{x_i, y_i\}_{i=1}^{\ell}$, сетка $(He)_{\ell=1}^{h}$, парам. η, λ. Ищем вектор весов всех связей $w = (w^1 \dots w^n)$

1) random $(x_i) \in X^\ell$; $\ell = 1 \dots h$, $n = 1 \dots He$

forward: $x_{in}^{\ell} = \sigma_n^{\ell}\left(\sum_{k=0}^{He_{\ell-1}} w_{kn}^{i} x_{in}^{\ell-1} \right)$

$z_{in}^{\ell} = (\sigma_n^{\ell})'\left(\sum_{k=0}^{He_{\ell-1}} w_{kn}^{i} x_{ik}^{\ell-1} \right)$ $\quad p_{n,i}^{h} = \frac{\partial h_i(w)}{\partial x_n^w}$

backward: $p_{ik}^{\ell-1} = \sum_{h=0}^{He_\ell} p_{in}^{\ell} z_{in}^{\ell} w_{kn}^{\ell}$

grad step for all $\ell = 1 \dots h$, $k = 0 \dots He_{\ell-1}$, $n = 1 \dots He_\ell$

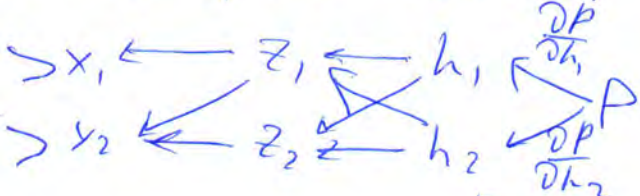
$w_{kn}^{\ell} = w_{kn}^{\ell} - \eta p_{in}^{\ell} z_{in}^{\ell} x_{in}^{\ell-1}$

Пока Q и/или весов изменяется

Пример:

③ $\frac{\partial P}{\partial h_1}$, $\frac{\partial P}{\partial h_2}$, ② $\frac{\partial P}{\partial z_i} = \frac{\partial P}{\partial h_1}\frac{\partial h_1}{\partial z_1} = \frac{\partial P}{\partial h_1}\frac{\partial h_2}{\partial z_1}$; $\frac{\partial P}{\partial z_2} = \frac{\partial P}{\partial h_1}\frac{dh_1}{\partial z_2} + \frac{\partial P}{\partial h_2}\frac{\partial h_2}{\partial z_2}$

① $\frac{\partial P}{\partial x_1} = \frac{\partial P}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial P}{\partial h_2}\frac{\partial h_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial P}{\partial h_1}\frac{\partial h_1}{\partial z_2}\frac{\partial z_2}{\partial x_1} + \frac{\partial P}{\partial h_2}\frac{\partial h_2}{\partial z_2}\frac{\partial z_2}{\partial x_1}$



Билет 4

Метод обучения Adaptive momentum. Метод Drop-out и его реализация

① $v_0 = 0$

$v_{k+1} = \beta_1 v_k + (1-\beta_1) \nabla f(x_k)$

$G_{k+1} = \beta_2 G_k + (1-\beta_2) (\nabla f(x_k))^2$

$x_{k+1} = x_k - \frac{e_2}{\sqrt{G_{k+1} + \varepsilon}} v_{k+1}$

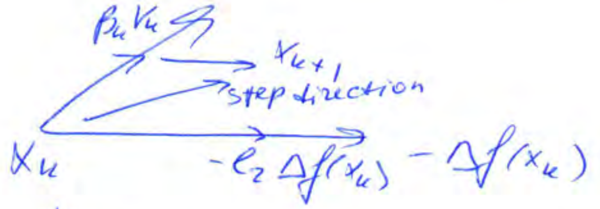
$\beta_1 = 0.9$

$\beta_2 = 0.999$

$\varepsilon = 1e-8$

$f(x) = \sum h(x, y_i)$

Получим из momentum и RMS prop. (Слер. 3)



RMS Prop: $G_{k+1} = \gamma G_k + (1-\gamma)(\nabla f(x_k))^2$

$x_{k+1} = x_k - \frac{l_2}{\sqrt{G_{k+1}+\varepsilon}} \nabla f(x_k)$

dropout

При град. шаге $h_i(w) \to \min_w$ отключаем n-ый нейрон l-го слоя с вер. p_l

$x^l_{n_i} = \xi^l_n \sigma^l_n(\sum w^l_{kn} x^{l-1}_{k_i})$ $P(\xi^l_n = 0) = p_l$

Вкл. с $(1-p_l)$

на обуч. $x^l_{n_i} = \frac{1}{1-p_l}\xi^l_n \sigma^l_n(\sum w^l_{kn} x^{l-1}_{k_i})$

на валидации $x^l_{n_i} = \sigma^l_n(\sum w^l_{kn} x^{l-1}_{k_i})$

с учетом h_2 gradstep $w = w(1-\eta\lambda) - \eta\frac{1}{1-p_l}\xi^l_n h_i'(w)$

Билет 5

Основные слои искусств. нейросетей: свертки, skip-connection, drop-out, batch norm, нелинейности.

1) Даны 2 матр. $A(n_x \times n_y)$ и $B(m_x \times m_y)$ $C = A * B$. C.shape = $(n_x - m_x + 1), (n_y - m_y + 1)$, т.е. $C_{i,j} = \sum_{u=0}^{m_x-1}\sum_{v=0}^{m_y-1} A_{i+u, j+v} B_{u,v}$

2) Skip-connection решает проблему затухания grad, передача инф. из нижних слоев в верхние:

skip-connection: → layer 1 → activation → layer 2 → ⊕ → ...

3) Dropout - случайное выключение некоторого числа нейронов для борьбы с переобучением и улучшения обобщающей способности от входн. данных

4) Batch norm - снимает зависимость $\mu = 0$, $\sigma = 1$ нелинейными слоями данные данные с

Batch := $\{x_1, \dots x_m\}$ $\mu_B = \frac{1}{m}\sum x_i$ $\sigma_B^2 = \frac{1}{m}\sum(x_i - \mu_B)^2$

$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}$ $y_i = \gamma \hat{x}_i + \beta = BatchNorm(x_i)$ - смещение и сдвиг.

5) нелинейность помогает моделировать сложные ф-ции.

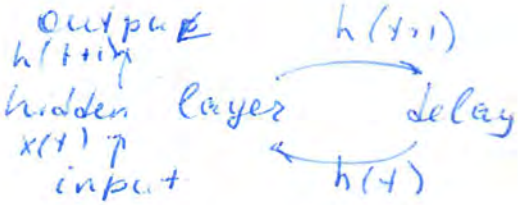
Билет 6

Рекуррентная нейронная сеть. LSTM.

RNN - сети с циклами, подходящие для обработки последовательностей. Для обучения применяется BPTT - Back propagation through time. Суть он такая:

Схема слоя RNN и развернутая схема.

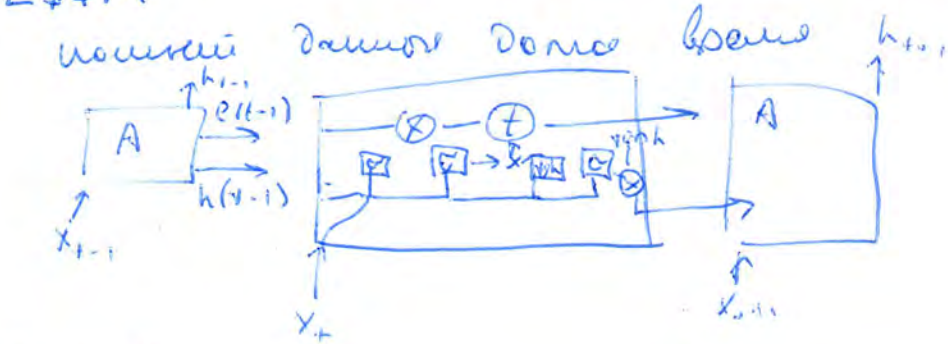
Схема с задержкой в скрытом слое:



Виды RNN:

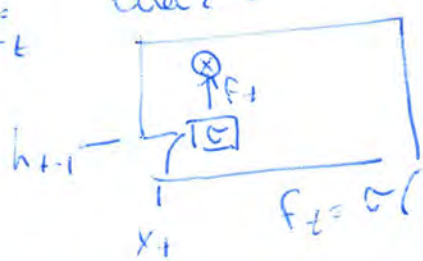
1) 1 вход, много выходов: для генерации аудио
2) Много входов 1 выход: для оценки тональностей
3) Много входов и выходов: для перевода

LSTM



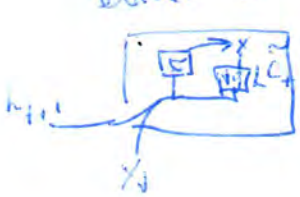
шаг 1

шаг 2



$f_t = \sigma(W_f(h_{t-1}, x_t) + b_f)$

шаг 3



шаг 4

$C_t = f_t C_{t-1} + i_t \tilde{C}_t$

$i_t = \sigma(W_i [h_{t-1}, x_t] + b_i)$

$\tilde{C}_t = tgh(W_t [h_{t-1}, x_t] + b_t)$

$h_t = \sigma(W_o [h_{t-1}, x_t] + b_o) tgh C_t$

Билет 7

Метод опорных векторов. Kernel trick.

Ставим задачу регрессии - дано train $X = \{(x_i, y_i)\}_{i=1}^{\ell}$
ищем $g(x)$ - спрогнозированную выборки путем обучения.
loss: $h(y, g(x)) = \begin{cases} 0, & |y - g(x)| \le \varepsilon \\ |y - g(x)| - \varepsilon, & |y - g(x)| > \varepsilon \end{cases}$, $\varepsilon > 0$
ищем решение в линейном виде $f(x) = (w, x) - w_0$
loss $a(x_i) = |(w, x_i) - w_0 - y_i|_\varepsilon \; \forall (x_i, y_i)$ где $|z|_\varepsilon = \max(0, |z| - \varepsilon)$
функционал, потерь
$Q_\varepsilon(a, X) = \sum_{i=1}^{\ell} |(w, x_i) - w_0 - y_i|_\varepsilon + \tau(w, w)^2 \to \min_{w, w_0}$
введем переменные ξ_i^+ и ξ_i^- - знач. которых = потерь при завышении или занижении ответа.
$\xi_i^+ = (a(x_i) - y_i - \varepsilon)_+ \quad \xi_i^- = (-a(x_i) + y_i - \varepsilon)_+ \quad i = 1, \dots \ell$
Задача минимизации:
$$\begin{cases} 0{,}5(w, w)^2 + \frac{1}{2\tau} \sum_{i=1}^{\ell} (\xi_i^+ + \xi_i^-) \to \min_{w, w_0, \xi_i^+, \xi_i^-} \\ (w, x_i) - w_0 \le y_i + \varepsilon + \xi_i^+ \\ (w, x_i) - w_0 \ge y_i - \varepsilon - \xi_i^- \\ \xi_i^- \ge 0, \; \xi_i^+ \ge 0 \end{cases}$$

Будем решать двойственную задачу, знач. прошлв. заменим одром $k(x_i, x_j)$, d_i^+, d_i^- - двойств. перем.
$$\begin{cases} h(d^+, d^-) = -\varepsilon \sum_{i=1}^{\ell} (d_i^+ + d_i^-) + \sum_{i=1}^{\ell} (d_i^- - d_i^+) y_i - \frac{1}{2} \sum (d_i^- - d_i^+)(d_j^- - d_j^+) \cdot k(x_i, x_j) \to \max_{d^+, d^-} \\ 0 \le d_i^+ \le C = \frac{1}{2\tau}, \quad 0 \le d_i^- \le C \\ \sum (d_i^+ + d_i^-) \to 0 \end{cases}$$
В рез. все x_i делятся на 5 типов
1) $|a(x_i) - y_i| < \varepsilon \quad d_i^+ = d_i^- = \xi_i^+ = \xi_i^- = 0$
2) $a(x_i) = y_i + \varepsilon \quad 0 < d_i^+ < C, \; d_i^- = \xi_i^+ = \xi_i^- = 0$
3) $a(x_i) = y_i - \varepsilon \quad 0 < d_i^- < C, \; d_i^+ = \xi_i^+ = \xi_i^- = 0$
4) $a(x_i) > y_i + \varepsilon \quad d_i^+ = C, \; d_i^- = 0, \; \xi_i^+ = a(x_i) - y_i - \varepsilon, \; \xi_i^- = 0$
5) $a(x_i) < y_i - \varepsilon \quad d_i^+ = 0, \; d_i^- = C, \; \xi_i^+ = 0, \; \xi_i^- = y_i - a(x_i) - \varepsilon$
7.5 - опорные, уд. в опр. всех. y-ки регрессии $\sum (d_i^- - d_i^+) k(x_i, x) - w_0$: $(w, x_i) - w_0 = \begin{cases} y + \varepsilon, & \xi_i \in 2 \\ y - \varepsilon, & \xi_i \in 3 \end{cases}$

Классификация:
$\varphi : \mathbb{R}^n \to H$, H - евкл. пр-во. сразу сведем задачу к лин. разд. выборке разд. ф-ция $f(x) = (w, \varphi(x)) + b$ где $w = \sum d_i y_i \varphi(x_i)$, где d_i зависит от y_i и $(\varphi(x_i), \varphi(x_j))$, ранее $k(x, y) = (\varphi(x), \varphi(y))$, k неотр. опр. и сим.
В случае лин. разд. выборки: ищем $f(x)$ т.ч. $f(x_i) > 0 \; \forall x_i \in \omega_1$, $f(x_i) < 0 \; \forall x \in \omega_2$, $y_i = \begin{cases} 1, & x \in \omega_1 \\ -1, & x \in \omega_2 \end{cases}$, т.е. $y_i((w, x_i) + b) \ge 0$
Разд. гиперплоскость $(w, x) + b = 0$, по т-ме Куна-Таккера задача опт. $h(w, b, \lambda) = 0{,}5(w, w) - \sum \lambda_i (y_i((w, x_i) + b) - 1) \to \min_{w, b} \max_{\lambda}$

$\lambda_i \geq 0$, $\lambda_i (y_i(\langle w, x_i\rangle + b) - 1) = 0$, отсюда $\lambda_i = 0$ или $y_i(\langle w, x_i\rangle + b) - 1 = 0$, т.е. x_i — опорный объект

Стр. 6

$$\begin{cases} \frac{\partial h}{\partial w_s} = w_s - \sum \lambda_i y_i x_{is} = 0 \\ \frac{\partial h}{\partial b} = -\sum \lambda_i y_i = 0 \end{cases} \Rightarrow w = \sum \lambda_i y_i x_i \text{ и } \sum \lambda_i y_i = 0$$

$L(w, b, \lambda) = \sum \lambda_i - 0{,}5 \sum \lambda_i \lambda_j y_i y_j (x_i, x_j) = \sum \lambda_i - 0{,}5 \| \sum \lambda_i y_i x_i \|^2$ т.е. квадрат и поэтому крит. имеет $\Phi(\lambda) = \sum \lambda_i - 0{,}5 \| \sum \lambda_i y_i x_i \|^2$

kernel trick

$w = \sum \lambda_i y_i \varphi(x_i)$ $\operatorname{argmax} f(\lambda_1, \dots \lambda_n) = \sum \lambda_i - \frac{1}{2} \sum_i \sum_j y_i \lambda_i k(x_i, x_j) y_j \lambda_j$

$\sum \lambda_i y_i = 0$, $0 \leq \lambda_i \leq \frac{1}{\ell \lambda}$ $\forall i$, $\varphi(x_i)$ легли на гран. в пр-ве

случая $b = w^T \varphi(x_i) - y_i = \sum y_j \lambda_j k(y_j, x_i) - y_i$; $z \mapsto \operatorname{sgn}(w^T \varphi(z) - b)$

Билет 8

Метод N-ближайших соседей. Оценка оптимальности классификации методом kNN (тер-во Кover - Харта)

Дан: train $X^m = \{(x_i, y_i)\}_{i=1}^m$, метрика $\rho(x, x')$ для произв. объекта, расположенных объектов x_i в порядке возр. до u $\rho(u, x_{1,u}) \leq \rho(u, x_{2,u}) \leq \dots \leq \rho(u, x_{m,u})$

$x_{i,u}$ — i-й сосед u тогда алгоритм будет задаваться ими

$a(u) = \operatorname{argmax}_{y \in Y} \sum [y_{i,u} = y] w(i, u)$ где $w(i, u)$ оценивает степень важности i-го соседа для объекта u. Для классического kNN $w(i, u) = [i \leq k]$ при таком $w(i, u)$ нар-во м.б. любым, тогда примем ядро Парзена: $w(i, u) = K\left(\frac{\rho(u, x_{i,u})}{h}\right)$ - парзеновское окно

N-во Кover - Харта: верхняя граница частота ошибок $R^* \leq R_{kNN} < 2R^*$. $R_{kNN} \leq 2R^* \left(1 - \frac{MR^*}{M-1}\right)$, R^* - частота ошибок BK R_{kNN} - асимпт. теория, M - число классов

Билет 9

Задача классификации. Основные метрики (confusion matrix, Rec, AUC) и их свойства. Оценка обобщ. способности модели.

~~Задача бин~~ Рассм. дан. бин. классификации $X \to Y$ $X = \mathbb{R}^n$, $Y = \{-1, +1\}$

$X^m = (x_i, y_i)_{i=1}^m$, имеет $a = \operatorname{sgn}(\langle w, x \rangle - w_0)$, ищем параметры w, w_0

рам. $K(X^m, w, w_0) = \sum [a \neq y_i] = \sum [(\langle w, x_i \rangle - w_0) y_i < 0]$

accuracy $= \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i] = \frac{TP + TN}{FP + FN + TP + TN}$

Матр. ошибок:

	pred +	pred −
+	TP	FN
−	FP	TN

При большом w_0 много x_i т.ч. $a(x_i) = -1$

добавл. ду-шарик, тогда $Loss = d_y [a(x_i) \neq y_i]$

ROC-AUC по x: $TPR = \frac{\sum [y_i = -1][a(x_i) = +1]}{\sum [y_i = -1]}$, по y: $TPR = \frac{\sum [y_i = +1][a(x_i) = +1]}{\sum [y_i = +1]}$

Оценка обобщ. способности: $X^\ell = \{(x_i, y_i)\}_{i=1}^\ell$

$y_i = y^*(x_i)$ $A_t = \{a, X \to Y\}$, $\mu_t = (X \times Y)^t \xrightarrow{\text{learning}} A_t$

$L(a, x)$ - Loss $Q(a, X^\ell) = \frac{1}{\ell} \sum L(a, x_i)$ = ф-ционал кач-ва

Самый частый: fold-ов $X^k = X_n^\ell \sqcup X_n^k$ $n = 1 \dots N$

cross-val $(\mu, X^\ell) = \frac{1}{|N|} \sum_{n \in N} Q_\mu(X_n^\ell, X_n^k)$