Пример: линейная регрессия

Линейный случай с m объектами и n признаками: $\mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w}$;
$\mathbf{y} \sim \mathcal{N}(\mathbf{f}(\mathbf{X}, \mathbf{w}), \beta^{-1}), \mathbf{w} \sim \mathcal{N}(0, \mathbf{A}^{-1})$.
Запишем интеграл:

$$p(\mathfrak{D}|\mathbf{h}) = p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \beta) = \frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\mathbf{w}} \exp\left(-0.5\beta(\mathbf{y} - \mathbf{f})^{\mathsf{T}}(\mathbf{y} - \mathbf{f})\right) \exp\left(-0.5\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w}\right) d\mathbf{w} =$$

$$= \frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\mathbf{w}} \exp(-S(\mathbf{w}))d\mathbf{w}$$

Для линейного случая интеграл вычисляется аналитически:

$$\int_{\mathbf{w}} \exp(-S(\mathbf{w}))d\mathbf{w} = (2\pi)^{\frac{n}{2}} \exp(-S(\hat{w}))|\mathbf{H}^{-1}|^{0.5},$$

где

$$\mathbf{H} = \mathbf{A} + \beta\mathbf{X}^{\mathsf{T}}\mathbf{X},$$

$$\hat{\mathbf{w}} = \beta\mathbf{H}^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

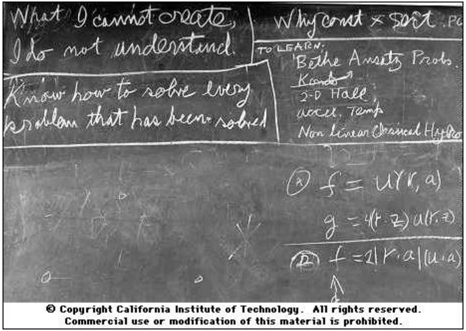
Вывод: для линейных моделей Evidence считается аналитически.

Порождающие и разделяющие модели

Московский Физико-Технический Институт

2021

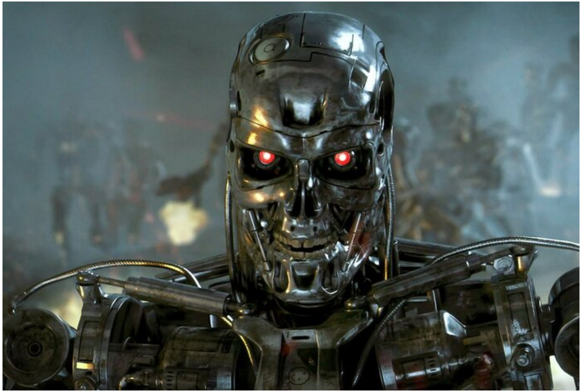
Идея порождающих (генеративных) моделей



Фейнман: *"Чего не могу воссоздать, того не понимаю."*

Идея разделяющих (дискриминативных) моделей

Платон: *"Человек — двуногое животное без перьев"*



Иногда проще решить целевую задачу (например, классификации), чем описать объект.

Генеративные и дискриминативные модели

Разделяющие модели

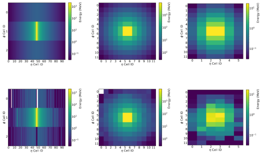
Моделируют: $p(y|\mathbf{x})$.

Порождающие модели

Моделируют: $p(y, \mathbf{x})$.

Порождающие модели:

- Порождение новых элементов выборки (когда генерация — самоцель)
- Создание синтетических данных для обучения/дообучения
- Получение скрытых свойств выборки (например, через латентные переменные)



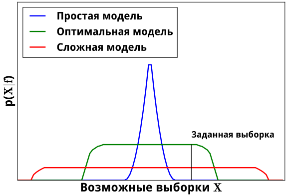
Выбор модели: связанный байесовский вывод

Первый уровень: выбираем оптимальные параметры:

$$\mathbf{w} = \arg\max \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathfrak{D}|\mathbf{h})},$$

Второй уровень: выбираем модель, доставляющую максимум обоснованности модели.
Обоснованность модели ("Evidence"):

$$p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$



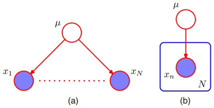
Что такое $\mathfrak{D}$ для генеративной и дискриминативной модели? И почему?

Plate notation

Plate notation — формат представления вероятностных моделей, альтернативный вероятностным графам.

Элементы:

- Белые кружки (случайные величины);
- Серые кружки (наблюдаемые реализации случайной величины);
- Маленькие кружки (неслучайные величины);
- Плитки (дублирование вероятностного вывода).



DAG и Plate notation (Bishop)

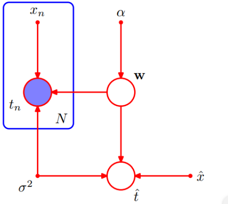
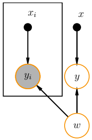


Plate notation для модели регрессии (Bishop)

Plate notation: дискриминативные и генеративные модели

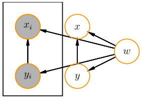
Дискриминативная модель:

- Порождается (или задан!) x
- Порождается w
- Порождается $Y \sim p(y|X, w)$



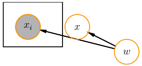
Генеративная модель:

- Порождается y
- Порождается w
- Порождается $x \sim p(X|y, w)$



Генеративная модель без учителя:

- Порождается w
- Порождается $x \sim p(X|w)$



Генеративные модели и обучение без учителя

Всегда ли генеративные модели обучаются без учителя?

Генеративные модели и обучение без учителя

Всегда ли генеративные модели обучаются без учителя?
Нет! Пример: линейные классификаторы

Логистическая регрессия:

$$\mathsf{E}\left(\boldsymbol{y}\,|\boldsymbol{X}\,\right) \equiv g^{-1}\left(\boldsymbol{X}\boldsymbol{w}\right),$$

$$g^{-1}(x)\frac{e^x}{1+e^x} \in [0,1]$$

Функция, задающая правило классификации, имеет форму сигмоиды.

Генеративная модель:

$$p(y=1|x,w) = \frac{p(x|w,y=1)p(y=1)}{\sum_{k=0}^{1} p(x|w,y=k)p(y=k)},$$

$$p(x|w,y=k) \sim \mathcal{N}(w_m^k, w_s^k).$$

Функция, задающая правило классификации, имеет форму сигмоиды.

Discriminative + generative

Наивный подход: введем априорное распределение на классы

$$p(\boldsymbol{x}, y|\boldsymbol{w}) = p(y|\boldsymbol{w}_y)p(x|y, \boldsymbol{w}_x).$$

Два вида оптимизации:

$$L_G = p(\boldsymbol{w}) \prod_{\boldsymbol{x},y} p(\boldsymbol{x}, y|\boldsymbol{w}),$$

$$L_D = p(\boldsymbol{w}) \prod_{\boldsymbol{x},y} p(y|\boldsymbol{x}, \boldsymbol{w}).$$

Можем комбинировать эти слагаемые:

$$\lambda L_G + (1-\lambda) L_D \to \max .$$

Оптимизация эвристична, не соответствует ни оптимизации ML, ни оптимизации MAP.

Discriminative + generative

(Bishop et al., 2007): введем две вероятностные модели: “дискриминативную” и “генеративную”:

$$p(\boldsymbol{x}, y | \boldsymbol{w}_G, \boldsymbol{w}_D) = p(y | \boldsymbol{x}, \boldsymbol{w}_D) p(\boldsymbol{x} | \boldsymbol{w}_G) p(\boldsymbol{w}_G, \boldsymbol{w}_D).$$

Максимизация:

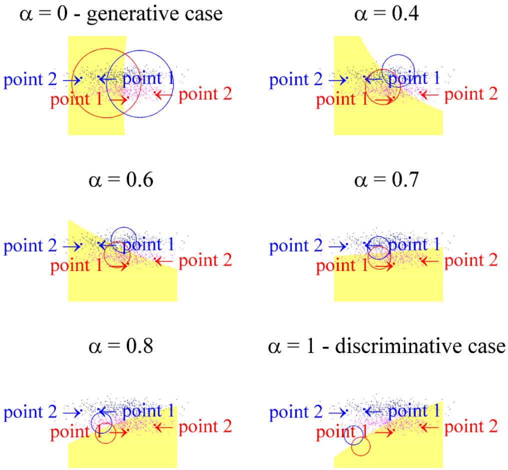
$$p(\boldsymbol{w}_G, \boldsymbol{w}_D) \prod_{\boldsymbol{x}, y} p(y | \boldsymbol{x}, \boldsymbol{w}_D) p(\boldsymbol{x} | \boldsymbol{w}_G).$$

Как назначить $p(\boldsymbol{w}_G, \boldsymbol{w}_D)$?

- $p(\boldsymbol{w}_G, \boldsymbol{w}_D) = p(\boldsymbol{w}_G) p(\boldsymbol{w}_D)$: получаем оптимизацию L_D;
- $p(\boldsymbol{w}_G, \boldsymbol{w}_D) = p(\boldsymbol{w}_G) \delta(\boldsymbol{w}_G - \boldsymbol{w}_D)$: получаем оптимизацию L_G;
- Компромисс: $p(\boldsymbol{w}_G, \boldsymbol{w}_D) \propto p(\boldsymbol{w}_G) p(\boldsymbol{w}_D) \mathrm{exp}(-\frac{1}{2\sigma^2} || \boldsymbol{w}_G - \boldsymbol{w}_D ||^2)$.

Discriminative + generative

(Bishop et al., 2007): пример разных комбинаций генеративной и дискриминативной модели для синтетической выборки. Размечено только по 2 точки каждого класса.



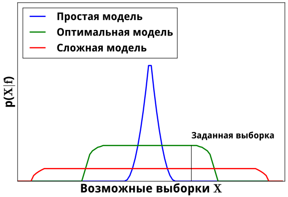
Задача выбора модели: revisited

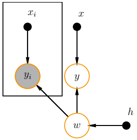
Первый уровень: выбираем оптимальные параметры:

$$\mathbf{w} = \arg\max \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathfrak{D}|\mathbf{h})},$$

Второй уровень: выбираем модель, доставляющую максимум обоснованности модели.
Обоснованность модели ("Evidence"):

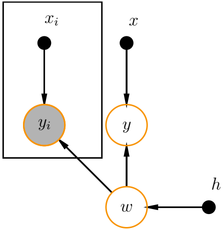
$$p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$

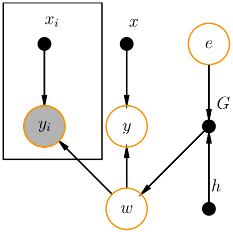




Задача выбора модели: revisited

Можно ли задать порождение параметров модели как еще одну модель?





Выбор модели: генеративный подход

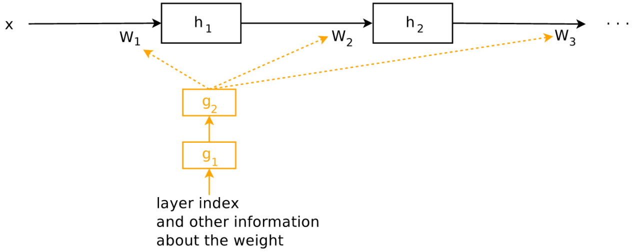
Определение

Пусть задано множество Λ.

Гиперсеть — это параметрическое отображение из множества Λ во множество параметров $\mathbb{R}^n$ модели $\boldsymbol{f}$:

$$G : \Lambda \times \mathbb{R}^u \to \mathbb{R}^n,$$

где $\mathbb{R}^u$ — множество параметров гиперсети.



Ha et al., 2016

Выбор модели: дискриминативный подход

$$\boldsymbol{w}_{\mathsf{MOE}} = \langle \gamma(\boldsymbol{x}), [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n] \rangle$$

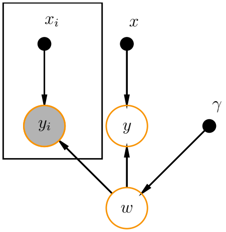


Рис. 1: Схема порождения выборки

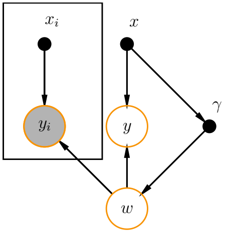


Рис. 2: Оптимизация МОЕ как дискриминативной модели

Литература и прочие ресурсы

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- Генератор котиков: https://github.com/aleju/cat-generator
- Paganini M., de Oliveira L., Nachman B. Accelerating science with generative adversarial networks: an application to 3D particle showers in multilayer calorimeters //Physical review letters. – 2018. – Т. 120. – №. 4. – С. 042003.
- Antoran J., Miguel A. Disentangling and learning robust representations with natural clustering //2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). – IEEE, 2019. – С. 694-699.
- Лекции по LDA: https://personal.utdallas.edu/~nrr150130/cs6347/2017sp/lects/Lecture_18_LDA.pdf
- Bernardo J. M. et al. Generative or discriminative? getting the best of both worlds //Bayesian statistics. – 2007. – Т. 8. – №. 3. – С. 3-24.
- Гребенькова О. С., Бахтеев О. Ю., Стрижов В. В. Вариационная оптимизация модели глубокого обучения с контролем сложности //Информатика и её применения. – 2021. – Т. 15. – №. 1. – С. 42-49.
- Ha D., Dai A., Le Q. V. Hypernetworks //arXiv preprint arXiv:1609.09106. – 2016.
- Адуенко А. А. Выбор мультимоделей в задачах классификации : дис. – Федер. исслед. центр"Информатика и управление"РАН, 2017.