BI-VZD přednáška 8

Alexander Kovalenko

FIT ČVUT

10. 4. 2022

Autoři: Karel Klouda, Juan Pablo Maldonado Lopez, Daniel Vašata.
Problémy, návrhy apod. hlaste v GitLabu.
Verze souboru: 11. dubna 2022 10:19.

Co bude v dnešní přednášce

- Představení klasifikace pomocí podmíněné pravděpodobnosti.
- Naivní Bayesův klasifikátor.
- Modely marginálních rozdělení v Naivním Bayesově klasifikátoru.
- Odhady parametrů těchto modelů včetně Baysovských odhadů.
- Představení rámce generativních $\times$ diskriminativních modelů.
- Využití Bayesova klasifikátoru ke klasifikaci textů.

"An individual has been described by a neighbour as follows: 'Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.' Is Steve more likely to be a librarian or a farmer?"

Daniel Kahneman (2011). Thinking, fast and slow.

"Steve je velmi plachý a uzavřený, vždy ochotný pomoci, ale má malý zájem o lidi nebo o svět reality. Je to tichá a spořádaná duše, má potřebu řádu a struktury a vášeň pro detaily. Bude Steve spíše knihovníkem, nebo farmářem?"

Daniel Kahneman (2011). Thinking, fast and slow.

Klasifikace na základě podmíněné pravděpodobnosti

Uvažujme nejprve klasifikační úlohu, ve které máme p diskrétních příznaků a chceme predikovat diskrétní vysvětlovanou proměnnou.

Místo přímé konstrukce prediktoru vysvětlované proměnné, např. pomocí náhodných lesů, zkusme postupovat následujícím pravděpodobnostním přístupem.

- Příznaky reprezentujeme náhodným vektorem $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ s hodnotami v $\mathcal{X}$ a vysvětlovanou proměnnou diskrétní náhodnou veličinou Y s oborem hodnot $\mathcal{Y}$.
- Na základě trénovací množiny odhadneme pravděpodobnosti $\mathrm{P}(Y = y|\boldsymbol{X} = \boldsymbol{x})$ pro každé $\boldsymbol{x} \in \mathcal{X}$ a $y \in \mathcal{Y}$.
- Tyto pravděpodobnosti můžeme využít k predikci Y při napozorovaných příznacích $\boldsymbol{x}$ takto:

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} \mathrm{P}(Y = y|\boldsymbol{X} = \boldsymbol{x}),$$

  tj. jako hodnotu, která je při daném $\boldsymbol{x}$ nejpravděpodobnější.
- Tato predikce se nazývá **MAP** odhad (z angl. **maximum a posteriori**).

Využití Bayesovy věty

- Otázkou je, jak odhadnout $\mathrm{P}(Y = y|\boldsymbol{X} = \boldsymbol{x})$.
- Mohli bychom se o to pokusit přímo na základě trénovacích dat. Tak se opravdu v mnoha případech postupuje (např. u neuronových sítí, logistické regrese, atd.).
- Nyní na to ale pojďme oklikou.
- Předpokládejme, že se nám podaří odhadnout $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)$.
- Využijeme-li Bayesovu větu z teorie pravděpodobnosti[1], můžeme kýženou pravděpodobnost spočíst jako

$$\mathrm{P}(Y = y|\boldsymbol{X} = \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)\,\mathrm{P}(Y = y)}{\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})},$$

kde

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}) = \sum_{y \in \mathcal{Y}} \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)\,\mathrm{P}(Y = y).$$

- Poznamenejme, že potřebujeme také odhad $\mathrm{P}(Y = y)$, který je ale triviální.

[1] Znáte z BI-PST!

Využití Bayesovy věty – predikce

- Zajímáme-li se o argument maxima tohoto výrazu vzhledem k y, můžeme zahodit jmenovatel $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})$, který bychom sice podle předchozího vztahu uměli spočítat, ale je pro všechny hodnoty y stejný.

- Tento fakt obvykle zapisujeme jako

$$\mathrm{P}(Y = y|\boldsymbol{X} = \boldsymbol{x}) \propto \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)\,\mathrm{P}(Y = y),$$

  kde symbolem $\propto$ myslíme rovnost, až na násobek konstantní vzhledem k y.

- Pro predikci tak finálně dostáváme

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)\,\mathrm{P}(Y = y).$$

- Nyní už zbývá „pouze“ odhadnout $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)$.

Naivní Bayes

Základem metody nazývané **naivní Bayes** nebo také **naivní Bayesův klasifikátor** (angl. **Naive Bayes**) je následující předpoklad:

Za podmínky $Y = y$ jsou všechny příznaky nezávislé.

Tj. pro každé $y \in \mathcal{Y}$ a $\boldsymbol{x} = (x_1, \ldots, x_p)^T \in \mathcal{X}$ platí

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x} | Y = y) = \mathrm{P}(X_1 = x_1 | Y = y) \cdot \ldots \cdot \mathrm{P}(X_p = x_p | Y = y).$$

Naivita tedy znamená, že pro fixní hodnotu vysvětlované proměnné předpokládáme, že jsou příznaky nezávislé.

Výsledný MAP odhad naivního Bayesova klasifikátoru je tedy

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} \prod_{i=1}^{p} \mathrm{P}(X_i = x_i | Y = y) \, \mathrm{P}(Y = y).$$

Ačkoliv je předpoklad podmíněné nezávislosti značně hrubý a povětšinou i nesprávný, dává naivní Bayes v mnoha případech až překvapivě dobré výsledky.

Naivní Bayes – příklad (1/2)

Uvažujme tři binární příznaky X_1, X_2, X_3 a binární vysvětlovanou proměnnou Y spolu s následující trénovací množinou:

Y	X_1	X_2	X_3
1	1	1	0
1	0	1	1
1	1	1	1
0	0	0	1
0	0	1	0
0	1	0	0

Nyní chceme provést predikci pro $\boldsymbol{x} = (0, 1, 1)^T$. Postupně dostáváme odhady[2]

$$\hat{p}(X_1 = 0|Y = 1) = \tfrac{1}{3}, \quad \hat{p}(X_2 = 1|Y = 1) = 1, \quad \hat{p}(X_3 = 1|Y = 1) = \tfrac{2}{3}.$$

Spolu s $\hat{p}(Y = 1) = \frac{1}{2}$ tak máme

$$\prod_{i=1}^{3} \hat{p}(X_i = x_i|Y = 1)\hat{p}(Y = 1) = \frac{1}{3} \cdot 1 \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{9}.$$

Analogicky platí

$$\prod_{i=1}^{3} \hat{p}(X_i = x_i|Y = 0)\hat{p}(Y = 0) = \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{27}.$$

Predikce je tedy $\hat{Y} = 1$, což je v tomto případě správně.

[2] Odhady pravděpodobností $\mathrm{P}(\dots)$ označíme $\hat{p}(\dots)$.

Naivní Bayes – příklad (2/2)

Nyní si řekněme, že ve skutečnosti jsou X_1, X_2, X_3 nezávislé veličiny se stejným rovnoměrným rozdělením, $X_i \sim \text{Be}(1/2)$, a $Y = 1$ právě, když jsou alespoň dvě hodnoty příznaků rovny 1.

Na základě pravděpodobností odhadnutých na trénovacích datech spočtěme predikce pro všechny možné body.

Y	X_1	X_2	X_3	$\hat{Y}$
1	1	1	0	1
1	0	1	1	1
1	1	1	1	1
0	0	0	1	0
0	0	1	0	0
0	1	0	0	0
1	1	0	1	0
0	0	0	0	0

Vidíme, že jediný případ, který je detekován špatně je $X_1 = 1, X_2 = 0, X_3 = 1$. To je docela dobrý výsledek, když vezmeme v potaz, že podmíněná rozdělení ve skutečnosti nejsou nezávislá. Platí totiž například

$$\mathrm{P}(X_1 = 0, X_2 = 0 | Y = 1) = \frac{\mathrm{P}(X_1 = 0, X_2 = 0, Y = 1)}{\mathrm{P}(Y = 1)} = \frac{\mathrm{P}(\emptyset)}{\mathrm{P}(Y = 1)} = \frac{0}{1/2} = 0$$

$$\mathrm{P}(X_1 = 0 | Y = 1) = \frac{\mathrm{P}(X_1 = 0, Y = 1)}{\mathrm{P}(Y = 1)} = \frac{\mathrm{P}(X_1 = 0, X_2 = 1, X_3 = 1)}{\mathrm{P}(Y = 1)} = \frac{1/2 \cdot 1/2 \cdot 1/2}{1/2} = \frac{1}{4}$$

a tudíž $\mathrm{P}(X_1 = 0, X_2 = 0 | Y = 1) = 0 \neq \frac{1}{16} = \mathrm{P}(X_1 = 0 | Y = 1) \cdot \mathrm{P}(X_2 = 0 | Y = 1)$.

Výsledek je o to zajímavější, uvědomíme-li si, že jsme na základě trénovací množiny získali špatné odhady pravděpodobností podmíněných hodnot příznaků.

Tato chyba se nejvíc projevila u bodu $\boldsymbol{x} = (1, 0, 1)^T$, kde jsme dostali $\hat{p}(X_2 = 0 | Y = 1) = 0$, což vedlo k nulovosti součinu odpovídajícího $Y = 1$ a následně k špatné predikci $\hat{Y} = 0$.

Diskuse vlastností naivního Bayese

- Navzdory faktu, že je předpoklad nezávislosti obvykle značně nepřesný, má naivní Bayesův klasifikátor několik dobrých vlastností.
- Především platí, že díky rozkladu sdružené podmíněné pravděpodobnosti $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)$ na součin marginálních podmíněných pravděpodobností $\mathrm{P}(X_i = x_i|Y = y)$ jsou vlastně příznaky separovány.
- Odhad podmíněných pravděpodobností každého příznaku tak probíhá nezávisle na ostatních.
- Tento fakt významně pomáhá rezistenci proti problémům s dimenzionalitou (the curse of dimensionality)!
- Jde totiž o to, že k rozumnému odhadu podmíněné pravděpodobnosti $\mathrm{P}(X_i = x_i|Y = y)$ nám stačí poměrně malé množství dat, a tento potřebný počet nenarůstá s nárustem počtu příznaků.

Diskuse vlastností naivního Bayese – pokračování

- Dále platí, že vzhledem k nepřesnému předpokladu nezávislosti obvykle nedostáváme dobrý odhad sdružené podmíněné pravděpodobnosti $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|Y = y)$.
- Naším cílem ale ve skutečnosti není tato pravděpodobnost jako taková, nýbrž MAP odhad, který z ní konstruujeme.
- Tento zůstane správný, pokud má skutečná hodnota y vyšší odhadnutou pravděpodobnost než ostatní hodnoty.
- Jak se ukazuje, v případě naivního Bayese je toto častá situace.
- Tento aspekt byl částečně patrný v předchozím příkladu, kde nezávislost neplatila a přesto byl kromě jediného případu výsledný odhad správný.

Modely podmíněných pravděpodobností – Bernoulliho rozdělení

- Zabývejme se nyní problematikou odhadu $\mathrm{P}(X = x|Y = y)$, kde X je jeden z příznaků.
- Nejjednodušší situace nastává pokud veličina X nabývá hodnot $0, 1$.
- V takovém případě je vhodným modelem pro podmíněné rozdělení $X|Y = y$ **Bernoulliho rozdělení** s parametrem p_y, tj. $\mathrm{P}(X = 1|Y = y) = p_y$. Značíme $(X|Y = y) \sim \mathrm{Be}(p_y)$.
- Jako odhad parametru p_y se nejčastěji využívá

$$\hat{p}_y = \frac{N_{1,y}}{N_{1,y} + N_{0,y}},$$

  kde $N_{1,y}$ značí počet dat pro $X = 1$ a $Y = y$, a $N_{0,y}$ analogicky.
- Z pohledu matematické statistiky se jedná o maximálně věrohodný odhad[3] (MLE odhad) parametru Beronulliho rozdělení.
- Tento odhad jsme použili v předchozím příkladu.

[3] Poznáte v BI-PST.

Modely podmíněných pravděpodobností – Bernoulliho rozdělení

- Nevýhodou tohoto odhadu je, že pro hodně malé (nebo velké) p_y se může stát, že v trénovací množině pro $Y = y$ nejsou zastoupeny obě hodnoty X_i a tedy $N_{1,y} = 0$ nebo $N_{0,y} = 0$.
- V takovém případě dojde ke kolapsu $\hat{p}_y = 0$ (nebo 1).
- Pokud se to stane, je odhad podmíněné pravděpodobnosti pro x_i rovné nevyskytující se hodnotě triviálně roven nule.
- MAP odhad $\hat{Y}$ pak bez ohledu na hodnoty ostatních příznaků nikdy nemůže predikovat příslušnou hodnotu y.
- Této situaci je možné se vyhnout Bayesovským přístupem s vhodným počátečním rozdělením.
- Pojďme si v krátkosti tento přístup představit.

Vsuvka – Bayesovský přístup k odhadům

- Uvažujme problematiku odhadu parametru p Bernoulliho rozdělení na základě naměřených dat $x_1, \ldots, x_n$.
- V klasickém frekventistickém přístupu statistiky konstruujeme odhad $\hat{p}$ pouze na základě napozorovaných dat.
- Do tohoto odhadu se nikterak nepromítá naše případná expertní znalost situace.
- Např. pokud odhadujeme pravděpodobnost panny p pro případ házení nějakou konkrétní existující mincí a padne nám 3 krát orel, odhadneme $\hat{p} = 0$.
- Tímto způsobem ale nejsme schopni nikterak zohlednit expertní posouzení situace, které zde může např.:

  Prostým pohledem a potěžkáním dané mince, se zdá, že bude standardní, a naše očekávání tedy je, že by p mělo být blízké $1/2$.

- **Bayesovská statistika** se s tímto problémem vypořádá zavedením takzvaného **apriorního rozdělení** (angl. **prior distribution**) odhadovaného parametru, která odráží naši expertní znalost, kterou hodláme na základě napozorovaných dat zpřesňovat.

Vsuvka – Bayesovský přístup k odhadům

- V našem příkladu tak předpokládáme, že parametr p má jako apriorní rozdělení spojité rozdělení na intervalu $(0, 1)$ určené hustotou $f_p(p)$, jejíž tvar určuje naše počáteční očekávání.
- Na základě napozorovaných dat $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ pak pomocí Bayesovy věty určíme takzvané **aposteriorní rozdělení** (angl.**posterior distribution**),

$$f_p(p|\boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|p) f_p(p)}{\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})},$$

  kde $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|p)$ je pravděpodobnost, že napozorujeme $\boldsymbol{X} = \boldsymbol{x}$ pokud p je ten správný parametr a

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}) = \int_p \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|p) f_p(p) \, \mathrm{d}p$$

  je vystředovaná pravděpodobnost, že napozorujeme $\boldsymbol{X} = \boldsymbol{x}$.
- Po využití pozorovaných dat tedy máme $f_p(p|\boldsymbol{x})$ aposteriorní rozdělení, které odpovídá změně našeho uvažování na základě pozorování.

Modely podmíněných pravděpodobností – Bernoulliho rozdělení

- Pokud na závěr potřebujeme získat bodový odhad parametru p (tj. jednu hodnotu), spočteme střední hodnotu aposteriorního rozdělení,

$$\hat{p} = \int_p p f_p(p|\boldsymbol{x})\,\mathrm{d}p.$$

- Chceme-li použít Bayesovský přístup v případně Bernoulliho rozdělení $X|Y = y$, musíme stanovit vhodné apriorní rozdělení.
- Obvykle se bere Beta rozdělení, jehož speciálním případem je rovnoměrné rozdělení (tj. apriorně předpokládáme, že všechny hodnoty p mohou být stejně pravděpodobné).
- Výsledný odhad z aposteriorního rozdělení je v tomto případě

$$\hat{p}_y = \frac{N_{1,y} + 1}{N_{1,y} + N_{0,y} + 2}.$$

- Tento odhad se nazývá **add-one smoothing** nebo také **Laplace's rule of succession** a jak je patrno, na kolaps $\hat{p}_y = 0$ netrpí.
- Implementace Bernoulliho rozdělení příznaků v knihovně `scikit-learn` je dostupná v balíčku `BernoulliNB`.

Modely podmíněných pravděpodobností – kategorické rozdělení

- Další možností je, když veličina X nabývá k různých hodnot $c_1, \ldots, c_k$.
- Vhodným modelem podmíněného rozdělení $X|Y = y$ je potom **kategorické rozdělení** (angl. také **multinoulli distribution**), které značíme $\mathrm{Cat}(\boldsymbol{p}_y)$, kde $\boldsymbol{p}_y = (p_{1,y}, \ldots, p_{k,y})^T$ a $p_{1,y}, \ldots, p_{k,y}$ jsou podmíněné pravděpodobnosti hodnot $c_1, \ldots c_k$, tj. $\mathrm{P}(X = c_j|Y = y) = p_{j,y}$.
- Jako odhad k-rozměrného parametru $\boldsymbol{p}_y = (p_{1,y}, \ldots, p_{k,y})^T$ se nejčastěji využívá

$$\hat{\boldsymbol{p}}_y = (\hat{p}_{1,y}, \ldots, \hat{p}_{k,y})^T \quad \text{a} \quad \hat{p}_{j,y} = \frac{N_{j,y}}{N_{1,y} + \ldots + N_{k,y}},$$

  kde $N_{j,y}$ značí počet dat pro $X = c_j$ a $Y = y$.
- I zde se jedná o maximálně věrohodný odhad, opět náchylný ke kolapsu, pokud se některá hodnota c_j v příslušné části trénovací množiny nevyskytuje.
- Analogicky je možné využít Bayesovský přístup a získat robustnější odhad

$$\hat{p}_{j,y} = \frac{N_{j,y} + 1}{N_{1,y} + \ldots + N_{k,y} + k}.$$

Modely podmíněných pravděpodobností – spojité rozdělení

- Uvažujme situaci, kdy je daný příznak X spojitou náhodnou veličinou.
- V takovém případě je $\mathrm{P}(X = x|Y = y) = 0$ pro každé x a použití předchozích vzorečků je bez vhodné modifikace nemožné.
- Abychom situaci zachránili, místo podmíněné pravděpodobnosti se pro tento příznak vezme podmíněná hustota pravděpodobnosti $f_{X|y}(x)$, což je hustota pravděpodobnosti veličiny X podmíněné jevem $Y = y$.
- Je to tedy hustota, která odpovídá distribuční funkci

$$F_{X|y}(x) = \mathrm{P}(X \leq x|Y = y).$$

- Predikci MAP odhadem provedeme s pomocí následujícího vztahu

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} \prod_{i=1}^{\ell} \mathrm{P}(X_i = x_i|Y = y) \prod_{i=\ell+1}^{p} f_{X_i|y}(x_i)\,\mathrm{P}(Y = y),$$

kde $X_1, \ldots, X_\ell$ jsou diskrétní příznaky a $X_{\ell+1}, \ldots, X_p$ jsou spojité příznaky.

Modely podmíněných pravděpodobností – Gaussovo rozdělení

- Častým modelem podmíněného rozdělení $X|Y = y$ je ve spojitém případě normální rozdělení $\mathrm{N}(\mu_y, \sigma_y^2)$ se střední hodnotou určenou parametrem μ_y a rozptylem určeným parametrem σ_y^2.
- Podmíněná hustota je tedy pro každé $x \in \mathbb{R}$ určena vztahem

$$f_{X|y}(x) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \mathrm{e}^{-\frac{1}{2\sigma_y^2}(x-\mu_y)^2}.$$

- Zde obvykle používáme MLE odhady

$$\hat{\mu}_y = \frac{1}{N_y}\sum_i^{N_y} x_i \quad \text{a} \quad \hat{\sigma}_y^2 = \frac{1}{N_y}\sum_i^{N_y}(x_i - \hat{\mu}_y)^2,$$

  kde $x_1, \ldots, x_{N_y}$ jsou hodnoty příznaku X, pro které $Y = y$.
- Implementace v knihovně `scikit-learn` je dostupná v balíčku `GaussianNB`.

Generativní přístup k predikci

- Většinu přednášky (kromě prvního slajdu) jsme se zabývali situací, kdy konstruujeme nějaký model pro odhad podmíněné pravděpodobnosti $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x} | Y = y)$.
- To nám spolu s odhadem $\mathrm{P}(Y = y)$ dává model pro sdruženou pravděpodobnost

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}, Y = y) = \mathrm{P}(\boldsymbol{X} = \boldsymbol{x} | Y = y)\, \mathrm{P}(Y = y).$$

- Tento přístup, kdy vytváříme model sdružené pravděpodobnosti $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}, Y = y)$, se obecně nazývá **generativní přístup** a výsledný model pak **generativní model**.
- Termín generativní znamená, že model sdružené pravděpodobnosti představuje úplnou informaci o rozdělení, ze kterého byla data „generována“.
- Generativní model tedy může být využit pro generování nových pozorování.
- Přeneseně se pak používá pojem generativní klasifikátor pro MAP odhad založený na $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}, Y = y)$.

Diskriminativní přístup k predikci

- Druhou možností je odhadovat na základě trénovacího datasetu podmíněnou pravděpodobnost $\mathrm{P}(Y = y|\boldsymbol{X} = \boldsymbol{x})$ napřímo, bez modelu pro sdruženou pravděpodobnost.
- Tomuto přístupu se říká **diskriminativní** a příslušný model podmíněné pravděpodobnosti $\mathrm{P}(Y = y|\boldsymbol{X} = \boldsymbol{x})$ se nazývá **diskriminativní model**.
- Analogicky se používá pojem diskriminativní klasifikátor pro výsledný MAP odhad založený na $\mathrm{P}(Y = y|\boldsymbol{X} = \boldsymbol{x})$.
- Jak uvidíme v následujících přednáškách, diskriminativní přístup je poměrně častý a většina v současné době nejpoužívanějších metod je tohoto typu.
- Jako příklad uveďme logistickou regresi nebo neuronové sítě.
- Někdy je jako diskriminativní přístup dokonce nazýván jakýkoliv přístup přímé predikce hodnot Y, tj. bez odhadu pravděpodobností.
- Do takové rozšířené definice pak spadají například i rozhodovací stromy.

Využití Bayesova klasifikátoru ke klasifikaci textů

- Jednou z reálných aplikací (naivního) Bayesova klasifikátoru je klasifikace textů na základě tzv. **bag-of-words** modelu.
- Při tomto přístupu je dokument reprezentován pomocí četností výskytů slov z nějakého slovníku $\mathcal{D}$.
- To znamená, že daný dokument má $D \equiv |\mathcal{D}|$ příznaků $X_1, \ldots, X_D$, kde X_j udává počet výskytů j-tého slova z $\mathcal{D}$.
- Nejjednodušší model pro podmíněnou pravděpodobnost je pak následující:

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x} | Y = y) = \frac{n!}{\prod_{j=1}^{D} x_j!} \prod_{j=1}^{D} p_{j,y}^{x_j},$$

  kde $p_{j,y}$ je pravděpodobnost, že náhodně vzaté slovo z dokumentu třídy y bude právě j-té slovo ze slovníku $\mathcal{D}$, a $n = \sum_j x_j$ je počet slov daného dokumentu.
- Při konstantní hodnotě n se výše uvedené rozdělení nazývá **multinomické rozdělení** a s parametry n a $p_{1,y}, \ldots, p_{D,y}$, kde $\sum_j p_{j,y} = 1$.

Využití Bayesova klasifikátoru ke klasifikaci textů

- Nyní je třeba na základě trénovací množiny N klasifikovaných dokumentů odhadnout parametry $p_{1,y}, \ldots, p_{D,y}$.
- Nejjednodušším odhadem je poměr počtu výskytů j-tého slova ve všech dokumentech dané kategorie dělený souhrnnou délkou všech dokumentů dané kategorie,

$$\hat{p}_{j,y} = \frac{N_{j,y}}{N_y},$$

  kde $N_{j,y} = \sum_{i=1}^{N} x_{i,j}$ přičemž $x_{i,j}$ je počet výskytů j-tého slova v i-tém dokumentu, a $N_y = \sum_{j=1}^{D} N_{j,y}$, je celkový počet slov ve všech dokumentech kategorie y.
- I zde je možné aplikovat Bayesovský přístup a získat add-one smoothing odhad

$$\hat{p}_{j,y} = \frac{N_{j,y} + 1}{N_y + D}.$$

- Implementace v knihovně `scikit-learn` je v balíčku `MultinomialNB`.