Vojtěch Pospíšil

DATA A INFORMACE, NÁHODA A PRAVDĚPODOBNOST, C# a SQL

Data a iformace, náhoda a pravděpodobnost, C# a SQL

Vojtěch Pospíšil

„Na počátku bylo slovo. A to slovo bylo u Boha…“, říká evangelista Jan.

A na zbývajících tisíci stranách hledá Bible pravdu.

Obojí neskončilo. Tvoření pokračuje. Vznikají nová slova přinášející nové pravdy (nebo staré lži).

Slovo informace stvořilo informatiku.

Slova logaritmus a pravděpodobnost vytvořila nástroj na měření informací.

Slovo počítač dalo vzniknout stroji na zpracování informací.

A slovo inflace vytvořilo nadbytek všeho – dat, informací, pravd i lží.

Tato kniha je o některých slovních zkratkách – C#, XAML, WPF, SQL.

A také snad trochu o nových informacích a starých pravdách.

Jak se liší data a informace?

Lze měřit informace pomocí náhody?

Jak poznat, které informace jsou pravdivé, a které lživé?

Proč je nevýhodné ukládat data do excelovské tabulky?

Kdy je lepší programovat přístup k databázi na konzolové platformě C#, a kdy na platformě Windows Presentation Foundation?

Které SQL dotazy do sebe vkládat, a které skládat?

Co mají společného opilost kapitána Browna a binární strom?

O tom všem je tato kniha. Na níže uvedeném odkazu najdete:

1) zdrojový kód v C jazyku
2) konzolové zdrojové kódy v jazyce C#
3) na platformě WPF použité kódy v jazycích C# a XAML
4) sadu SQL dotazů použitých při zobrazování informací z knihovní databáze
5) naplněnou knihovní databázi

https://eknihyjedou.cz/content/knihovnaDb.zip

Kapitola 1:

Data a informace, náhoda a pravděpodobnost. Jak veliká je informace?

Data jsou údaje, které získáváme pozorováním nebo měřením. Interpretací vztahů mezi daty vznikají **informace**. To, co spouští interpretaci vztahů mezi daty, je **otázka**.

Chceme-li informace měřit, musíme mít stanovenu nějakou nejmenší jednotku, na kterou budeme velikost informace přepočítávat. Je-li otázka spouštěčem hledání informací, pak **odpověď** je výsledkem tohoto hledání. Existuje nějaká univerzální nejstručnější odpověď? V češtině je to odpověď typu ANO/NE, v počítačovštině 1/0. Dvojice 1 nebo 0 se anglicky nazývá **binary digit**, zkráceně bit, což je jednotka informace. Jeden bit je tedy číslice, která má hodnotu 1 nebo 0.

Chceme-li změřit velikost výsledné informace, musíme k ní dojít sérií takových otázek, na které lze odpovídat pouze ano nebo ne. Počet otázek určuje počet bitů, čili velikost výsledné informace. Otázky musíme volit tak, aby jich bylo co nejmenší nutné množství. Kolik jich bude? Umět se správně a stručně ptát je pro mnoho lidí problémem. Neobratný tazatel se požadovanou informaci doví po 20 otázkách, logicky postupující tazatel třeba jen po 6 otázkách. Dojdeme tak ke dvěma různým velikostem téže informace. Jak měření informací učinit jednoznačným a nezpochybnitelným? Pomůže nám matematika. Konkrétně teorie pravděpodobnosti.

Exkurze do teorie pravděpodobnosti:

Všechno, co je měřitelné, má jako výsledek měření přiřazeno číselnou hodnotu. Ale i jakémukoliv ještě nenastalému jevu dokáže matematika předem přiřadit číselnou hodnotu. Jevu, který určitě nastane, přiřazuje hodnotu 1 (jev jistý). Jevu, který určitě nenastane, přiřazuje hodnotu 0 (jev nemožný). A všechna desetinná čísla mezi nulou a jedničkou představují menší či větší pravděpodobnost budoucího nastání nějakého uvažovaného jevu. K určení pravděpodobnosti jednoduchého jevu slouží klasická definice pravděpodobnosti. Pravděpodobnost je rovna počtu možností zkoumanému jevu příznivých dělenému počtem všech možností.

Př. Jaká je pravděpodobnost, že při hodu mincí padne líc? P=1/2=0,5

Př. Jaká je pravděpodobnost, že při hodu kostkou padne pětka? P=1/6=0,1$\overline{6}$

Př. Jaká je pravděpodobnost, že při hodu kostkou padne sudé číslo? P=3/6=0,5

Ptáme-li se na pravděpodobnost složeného jevu, skládajícího se z několika jevů jednoduchých, získáme výslednou pravděpodobnost z pravděpodobností jevů jednoduchých tak, že je spojujeme buďto násobením nebo sčítáním.

Př. Jaká je pravděpodobnost, že při hodu dvěma kostkami padnou dvě šestky? $P = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = 0{,}02\overline{7}$

Př. Jaká je pravděpodobnost, že při hodu dvěma kostkami padnou dvě šestky nebo dvě pětky? $P = \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{2}{36} = 0{,}0\overline{5}$

Pokud zkoumáme pravděpodobnost složeného jevu, kde pořadí nastání jevů jednoduchých je předem v otázce dáno, postupujeme tak, jako dosud.

Př. Jaká je pravděpodobnost, že při třikrát opakovaném hodu kostkou padne v prvním a třetím hodu šestka? $P = \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \frac{5}{216} = 0{,}023\overline{148}$

Pokud ovšem pořadí nastání jednoduchých jevů dáno není, musíme pravděpodobnost jednoho pořadí násobit počtem všech možných pořadí. Tento počet je roven permutaci s opakováním.

Př. Jaká je pravděpodobnost, že při třikrát opakovaném hodu kostkou padnou dvě šestky? $P = \frac{3!}{2! \cdot 1!} \cdot 0{,}023\overline{148} = 3 \cdot 0{,}023\overline{148} = 0{,}069\overline{4}$

Př. Jaká je pravděpodobnost, že při třikrát opakovaném hodu kostkou padnou alespoň dvě šestky? $P = 3 \cdot 0{,}023\overline{148} + \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = 0{,}069\overline{4} + 0{,}004\overline{629} = 0{,}\overline{074}$

K úspěšnému zvládnutí pravděpodobnostních výpočtů nám chybí poslední nástroj – kombinační číslo dostupné na všech kalkulačkách jako nCr, ale v matematickém zápisu jako $\binom{n}{r}$.

Př. Jaká je pravděpodobnost, že z osudí obsahujícího 4 bílé a 5 černých koulí vylosuju 4 koule tak, že nejméně 3 budou černé?

a) Buďto vycházím z toho, že koule táhnu postupně: P = $\frac{4!}{3!\cdot 1!}\cdot\frac{5}{9}\cdot\frac{4}{8}\cdot\frac{3}{7}\cdot\frac{4}{6}+\frac{5}{9}\cdot\frac{4}{8}\cdot\frac{3}{7}\cdot\frac{2}{6}=\frac{20}{63}+\frac{5}{126}=\frac{45}{126}=0{,}3\overline{571428}$ … podmíněná pravděpodobnost

b) Nebo koule vyberu jediným tahem a použiju kombinační čísla: P = $\frac{\binom{5}{3}\cdot\binom{4}{1}}{\binom{9}{4}}+\frac{\binom{5}{4}\cdot\binom{4}{0}}{\binom{9}{4}}=\frac{20}{63}+\frac{5}{126}=\frac{45}{126}=0{,}3\overline{571428}$ … klasická pravděpodobnost

Obdržená informace bude tím větší, čím méně bude očekávaná (pravděpodobná). Velikost informace = - $\log_2 P(i)$ bitů.

Informace, že na minci padl líc má hodnotu $-\log_2 0{,}5 = 1$ bit.

Informace, že na kostce padla šestka má hodnotu $-\log_2\frac{1}{6}$ = 2,584962501 bitů.

Informace, že ze tří hodů kostkou padly dvě šestky má hodnotu $-\log_2\frac{15}{216}$ = 3,847996907 bitů.

Jeden bit je takové množství informace, které odstraňuje míru neurčitosti mezi dvěma stejně pravděpodobnými náhodnými jevy.

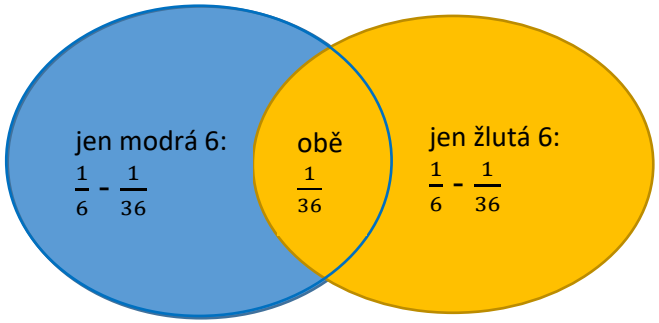
Vysvětlení účelu použití teorie pravděpodobnosti v informatice bylo touto definicí splněno.

Základní softwarový nástroj programátora pracujícího s pravděpodobností je **generátor pseudonáhodných čísel**. Dokáže totiž modelovat řadu problémů násobně rychleji než hledání analytických vzorců. Ukážeme si to na následujících stránkách.

Teorie pravděpodobnosti je v řadě matematických disciplín jediná, u které známe přesné datum jejího vzniku. Rytíř Chevalier de Méré se kromě poezie, milostných románků a hudby věnoval na dvoře Krále Slunce Ludvíka XIV i hře v kostky. Královské dvořany obehrával nalákáním na sázku: „*Hodím 4x kostkou a vsadím si proti vám, že mi alespoň jednou padne šestka*." Když vystřídal všechny hráče, vymyslel novou sázku: „*Hodím 24x dvěma kostkami a vsaďte se, že mi alespoň 1x padnou dvě šestky*." Nový sázkařský model ale nefungoval. Místo výher následovaly prohry. Obrátil se proto na mnohem mladšího Blaise Pascala (známého tvůrce prvního fungujícího a „sériově vyráběného" počítacího stroje Pascalina), se kterým se sblížil v roce 1653 na výletě do kraje Poitou (odkud pocházel, konkrétně z města Poitiers). „V čem je chyba? Hodím-li 1x kostkou, pravděpodobnost, že mi padne šestka, je jedna šestina. Hodím-li 2x kostkou, je pravděpodobnost, že mi alespoň jednou padne šestka přirozeně větší. A při čtyřech hodech kostkou je tedy 1/6+1/6+1/6+1/6=4/6=2/3. Proto víc vyhrávám, než prohrávám. Změním-li model na 24 hodů dvěma kostkami a pravděpodobnost, že mi alespoň jednou padnou dvě šestky, pak výsledná pravděpodobnost je přeci stejná (2/3). Při hodu dvěma kostkami je pravděpodobnost padnutí dvou šestek 1/36 (všech možných dvojic výsledků je totiž 36 – 1:1,1:2, 1:3, 1:4, 1:5, 1:6, 2:1, 2:2, atd… až 6:5, 6:6). Při 24 hodech dvěma kostkami je výsledná pravděpodobnost rovna 1/36 +1/36 + …1/36 = 24*1/36=2/3."

Blaise Pascal se s tímto problémem obrátil v srpnu 1654 (datum vzniku teorie pravděpodobnosti) dopisem na Pierra Fermata, který byl stejně jako jeho otec Etienne (v té době již mrtvý) právník, a matematik amatér. Oba společně dospěli k závěru, že rytíř de Méré měl štěstí již při první sázce, neboť jeho výpočet byl chybný.

Ukážeme si chybu na jednodušším případě: „*Hodím-li 2x kostkou, jaká je pravděpodobnost, že mi alespoň 1x padne šestka?*" Protože je jedno, jestli hodím 2x po sobě jednou kostkou nebo 1x dvěma kostkami, vyrobíme si množinový obrázek pro druhou variantu. Jedna kostka bude modrá, druhá žlutá. Zadání příkladu vyhovují 3 výsledky. Buď padne šestka jen na modré kostce nebo jen na žluté kostce nebo na obou kostkách současně. Pravděpodobnost padnutí šestky jen na modré kostce je $\frac{1}{6} \cdot \frac{1}{36}$, pravděpodobnost padnutí šestky jen na žluté kostce je $\frac{1}{6} \cdot \frac{1}{36}$ a pravděpodobnost padnutí šestky na obou kostkách současně (**věta o násobení pravděpodobnosti**) je $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. Řešení naší pravděpodobnosti odpovídá číslo $\frac{1}{6} \cdot \frac{1}{36} + \frac{1}{6} \cdot \frac{1}{36} + \frac{1}{36} = \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{36} = \frac{11}{36}$.



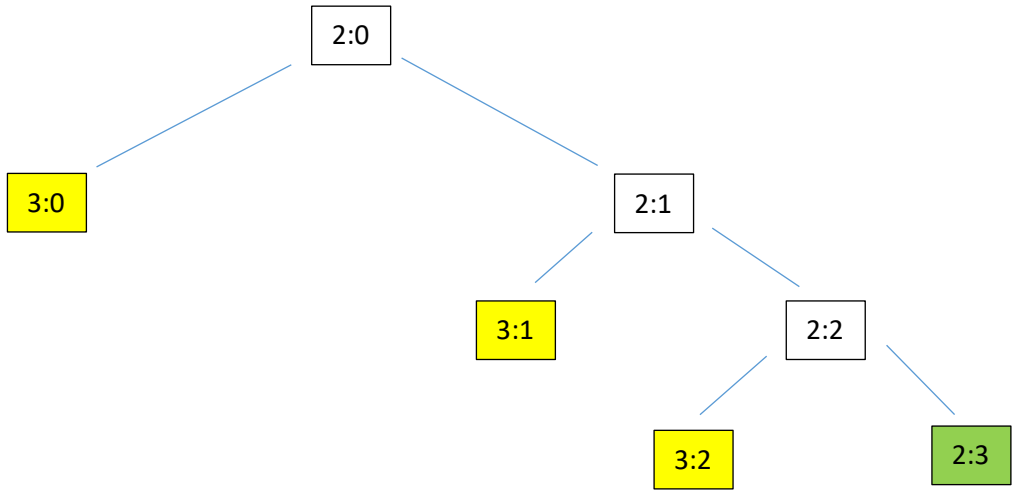
Množinové řešení pro 4 kostky by bylo velmi komplikované. Fermat s Pascalem proto našli elegantnější postup pomocí **doplňkové pravděpodobnosti**. Je jisto (pravděpodobnost 1), že určitě nastane jeden ze 4 případů: šestka na obou kostkách nebo šestka jen na modré nebo šestka jen na žluté nebo šestka ani na jedné kostce: $\frac{1}{36} + \frac{1}{6} - \frac{1}{36} + \frac{1}{6} - \frac{1}{36} + \frac{5}{6} \cdot \frac{5}{6} = 1$. Po úpravě: $\frac{1}{6} + \frac{1}{6} - \frac{1}{36} = 1 - \frac{25}{36}$.

Řešení úlohy rytíře de Méré je: $1 - \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} = 1 - \frac{625}{1296} = \frac{671}{1296} = 0{,}5177$. Srovnání s chybným výpočtem: $\frac{2}{3} = 0{,}6666$.

Pravděpodobnost padnutí minimálně jedné dvojice šestek z 24 hodů je: $1 - (\frac{35}{36})^{24} = 0{,}4914$.

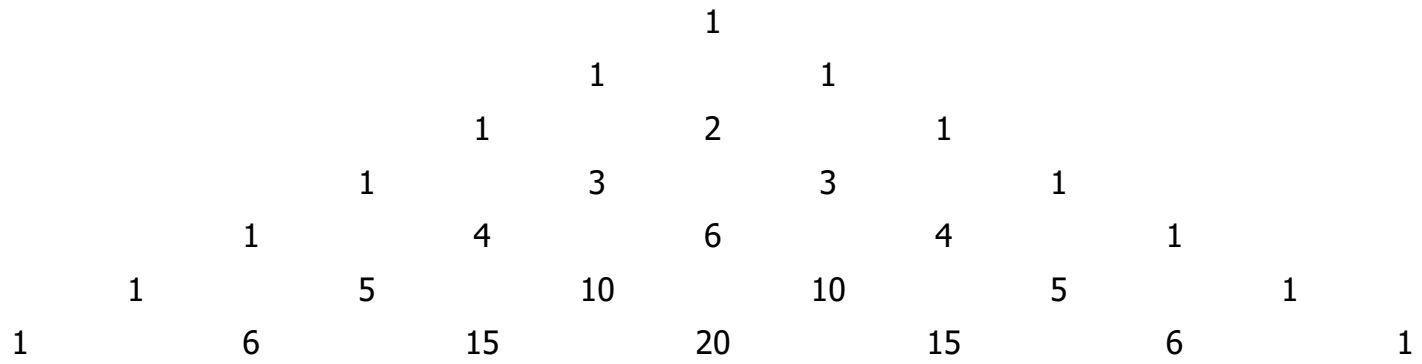
Pro „velký úspěch" rytíř Pascala požádal o řešení ještě jednoho hráčského problému. *„Hraji s protihráčem hru na 3 vítězství. Když vedu už 2:0 a zbývá mi k zisku vsazeného obnosu jediná výhra, král zavelí konec a máme dohráno. Jak si spravedlivě rozdělíme bank? Protihráč chce vrátit sázky, já si chci vzít celý bank, protože jsem už skoro vyhrál."*

Fermat s Pascalem použili k vyřešení graf, který dnes zná každý informatik: **binární strom**.



Každá větev stromu nastane s pravděpodobností $\frac{1}{2}$. Pravděpodobnost vítězství prvního hráče za stavu 2 : 0 spočítáme jako $P_1 = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{7}{8}$. Pravděpodobnost vítězství druhého hráče za stavu 2 : 0 je $P_2 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$. Spravedlivé rozdělení banku je tedy 7 ku 1.

Nebyl by to ale Pascal, aby nehledal jednodušší řešení než kreslení binárního stromu. A zde se na scéně dějin ocitá poprvé „Pascalův trojúhelník".



Řešení předchozího příkladu pomocí Pascalova trojúhelníku je jednoduché. Prvnímu hráči chybí k vítězství 1 hra, druhému 3 hry. Součet chybějících her pro prvního a pro druhého hráče je 4. Najdeme proto v trojúhelníku řádek, který obsahuje 4 čísla. Součet kombinačních čísel tohoto řádku je 8. Na tolik dílů se bude dělit bank. Tři hry představují součet prvních tří čísel: 1 + 3 +3 = 7. Jedna hra je číslo poslední, tedy 1. Bank se tedy rozdělí v poměru 7 ku 1.

Představme si jiný příklad. Hru na 5 vítězství. Hra je za stavu 3 : 2 přerušena. Jak si hráči rozdělí bank? K vítězství chybí prvnímu hráči 2 hry, druhému 3 hry. Součet je tedy 5, čemuž odpovídá pátý řádek trojúhelníka. Součet čísel řádku je 1 + 4 + 6 + 4 + 1 = 16. Bank se rozdělí na 16 dílů, které se budou dělit v poměru 11 (1+4+6) ku 5 (4+1).

Opusťme teď „báječný" dvůr Krále Slunce Ludvíka XIV (po něm vládl až jeho pravnuk Ludvík XV, protože zbývající část rodiny mu při „královských" hostinách postupně otrávili) a pojďme do nábřežní krčmy protestantského přístavního města La Rochelle. Tři mušketýry ani kardinála Richelieua tam už bohužel nepotkáme, zato *se od dveří přístavní krčmy snaží dojít ke své lodi opilý kapitán Brown. Naštěstí loď kotví jen 8 kroků od krčmy a podél nábřeží vede zábradlí, jehož se kapitán přidržuje. Ale protože pravděpodobnost kroku vpřed je při opilosti kapitána Browna stejná jako pravděpodobnost kroku vzad, je otázkou s jak velkou pravděpodobností se mu podaří po maximálně 20 krocích (na víc již kapitán Brown nemá sil) dojít k lodi.*

Řešme analyticky nejprve jednodušší případ, kdy loď kotví jen 2 kroky daleko, a hledejme pravděpodobnost, že kapitán dojde ke své lodi po 2, 4, 6, … krocích. Jako nástroj řešení použijeme opět binární strom.