Data Bases Data Mining
Functional dependencies exercise sheet

We recall the following inference rules, where reflexivity, augmentation and transitivity constitute the so called Armstrong's system.

$$\frac{Y \subseteq X}{X \to Y}\ \sigma_R \text{ (reflexivity)} \qquad \frac{X \to Y \qquad X \to Z}{X \to YZ}\ \sigma_C \text{ (composition)}$$

$$\frac{X \to Y}{WX \to WY}\ \sigma_A \text{ (augmentation)} \qquad \frac{X \to YZ}{X \to Y}\ \sigma_D \text{ (decomposition)}$$

$$\frac{X \to Y \qquad Y \to Z}{X \to Z}\ \sigma_T \text{ (transitivity)} \qquad \frac{X \to Y \qquad WY \to Z}{WX \to Z}\ \sigma_P \text{ (pseudo-transitivity)}$$

A proof of $X \to Y$ from Σ written $\Sigma \vdash X \to Y$ is a *sequence* $\langle f_0, \dots, f_p \rangle$ of FDs s.t. $f_p = X \to Y$ and $\forall i \in [0..p]$ either $f_i \in \Sigma$, or f_i is the *conclusion* of a rule from the Armstrong's system of which all its *antecedents* $f_0 \dots f_p$ appear before f_i in the sequence.

Exercice 1 : Proof of Functional Dependencies (FD)

Let Σ be the following set of FDs:

$$\begin{array}{ll} BC \to A & D \to BE \\ AC \to B & B \to DE \\ AE \to C & C \to E \end{array}$$

1. Prove using Armstrong's system that the following dependencies are entailed by Σ
    1. $AD \to C$
    2. $AB \to C$
    3. $AE \to BD$
    4. $AC \to D$
    5. $CD \to A$

Exercice 2 : Inference rules for FDs

1. Is the following inference rule correct? If true, prove it, otherwise, exhibit a counter-example.
$$\frac{XW \to Y \qquad XY \to Z}{X \to (Z \setminus W)}$$
2. Show that any proof $F \vdash X \to Y$ using the σ_P rule can be transformed into a proof that uses solely σ_A et σ_T.
3. Show that any proof $F \vdash X \to Y$ using the σ_R, σ_A and σ_T rules can be transformed into a proof that uses solely σ_R et σ_P.
4. Conclude that the set of rules $\{\sigma_R, \sigma_P\}$ is sound and complete for the inference problem of FDs.

Exercice 3 : The syntatic closure is a closure

The *syntactic closure* of X w.r.t. Σ is defined as $X^\star = \{A \mid \Sigma \vdash X \to A\}$. In the usual (algebraic) sense a *closure* is a mapping $\phi : \wp(E) \to \wp(E)$[1] that satisfies the three following properties:

Extensive $X \subseteq \phi(X)$

Increasing $X \subseteq Y \Rightarrow \phi(X) \subseteq \phi(Y)$

Idempotent $\phi(\phi(X)) = \phi(X)$

1. Show that $X^\star$ is a closure in the algebraic sense using the Armstrong's system. For idempotency, you may prove first that $\Sigma \vdash X \to X^\star$.
2. Show that for all set of attributes Y, if $X \subseteq Y \subseteq X^\star$ then $Y^\star = X^\star$
3. Use the previous property to compute the *set of closed sets* $CI(\Sigma) = \{X^\star \mid X \subseteq R\}$ for Σ as defined in exercise 1.

Exercice 4 : Functional dependencies and propositional logic (*[1]*)

The goal of this exercise is to relate functional dependencies and (classical) propositional logic, to highlight the point that the syntactic resemblance between a FD and logical implication is not a sheer one.

Let $\mathcal{P}$ be an infinite enumerable set of propositional variables and let Σ be a set of FD over $\mathbf{R}$. For each attribute $A \in \mathbf{R}$ we associate a corresponding *propositional variable* $\underline{A} \in \mathcal{P}$. This mapping is extended to FD: for each FD $A_1 \dots A_p \to B_1 \dots B_q \in \Sigma$ we associate the propositional formula $\underline{A_1} \wedge \dots \wedge \underline{A_p} \Rightarrow \underline{B_1} \wedge \dots \wedge \underline{B_q}$ with variables in $\mathcal{P}$. Finally, we write $\underline{\Sigma}$ for the set of propositional formulas associated to Σ. Let $\alpha = A_1 \dots A_p \to B_1 \dots B_q$ be a FD, we want to show that the following are equivalent:

$$\Sigma \models \alpha \tag{1}$$
$$\Sigma \models_2 \alpha \tag{2}$$
$$\underline{\Sigma} \models_P \underline{\alpha} \tag{3}$$

We write $\Sigma \models \alpha$ for the usual logical entailment between a set of FD Σ and a single FD α. We write $\Sigma \models_2 \alpha$ for the logical entailment restricted to the case where *relations have at most two tuples*. In other words, $\Sigma \models_2 \alpha$ is defined as $\forall r.(|r| = 2 \wedge r \models \Sigma) \Rightarrow r \models \alpha$. Finally, $\underline{\Sigma} \models_P \underline{\alpha}$ is the classical propositional logical entailment, i.e., for all assignment of propositional variables $\nu : \mathcal{P} \to \{0, 1\}$ such that $\nu \models \underline{\Sigma}$ it is the case that $\nu \models \underline{\alpha}$ as well.

Lemma 1. *Let $\nu : \mathcal{P} \to \{0, 1\}$ be an assignment of propositional variables, and let $r = \{t_1, t_2\}$ be the instance with two tuples t_1 and t_2 such that $t_1[A] = 1$ and $t_2[A] = \nu(\underline{A})$ for all $A \in \mathbf{R}$. Let $\alpha = A_1 \dots A_p \to B_1 \dots B_q$ be a FD. Then $\nu \models_P \underline{\alpha}$ if and only if $r \models \alpha$.*

1. Show that proposition (1) is equivalent to proposition (2). For the (2) $\Rightarrow$ (1) direction, use proof by contradiction (show that $\Sigma \models_2 \alpha$ and $\Sigma \not\models \alpha$ is inconsistent).
2. Show lemma 1 by constructing a two tuples instance "à la Armstrong" from ν and vice-versa.
3. Show that proposition (3) is equivlente to proposition (2) using lemma 1. Use proof by contradiction for each direction.
4. Conclude the main theorem.

References

[1] R. Fagin. Functional dependencies in a relational database and propositional logic. *IBM J. Res. Dev.*, 21(6):534–544, Nov. 1977.

[1] $\wp(E)$ is the set of all subsets of E, formally $\wp(E) = \{X \mid X \subseteq E\}$

Corrections

Solution de l'exercice 1

1. 1. $$\dfrac{\dfrac{\dfrac{D \to BE}{AD \to ABE}\,\text{aug.} \qquad \dfrac{AE \subseteq ABE}{ABE \to AE}\,\text{refl.}}{AD \to AE}\,\text{trans.} \qquad AE \to C}{AD \to C}\,\text{trans.}$$

2. $AB \to C$
   $B \to DE$ donc $AB \to ADE$ par augmentation
   $AB \to AE$ par décomposition
   $AB \to C$ puisque $AE \to C$ par transitivité

3. $AE \to BD$
   $AE \to C$ donc $AE \to AC$ par augmentation
   $AC \to B$ donc $AE \to B$ par transitivité
   $B \to DE$ donc $B \to D$ par décomposition
   On en déduit $AE \to D$ par transitivité
   Par composition, on a $AE \to BD$.

4. $AC \to D$
   $B \to DE$ donc $B \to D$ par décomposition
   Or $AC \to B$ donc donc $AC \to D$ par transitivité

5. $$\dfrac{\dfrac{\dfrac{D \to BE}{D \to B}\,\text{decomp.}}{CD \to BC}\,\text{aug.} \qquad BC \to A}{CD \to A}\,\text{trans.}$$

Solution de l'exercice 2

1. Non, on exhibe un contre-exemple avec l'instance suivante qui satisfait $XW \to Y$ et $XY \to Z$ mais pas $X \to (Z \setminus W)$:

W	X	Y	Z
w_0	x_0	y_0	z_0
w_1	x_0	y_1	z_1

2. Il s'agit de montrer que l'on peut prouver $WX \to Z$ à partir de $X \to Y$ et $WY \to Z$ en utilisant uniquement σ_A et σ_T :

$$\dfrac{\dfrac{X \to Y}{WX \to WY}\,\sigma_A \qquad WY \to Z}{WX \to Z}\,\sigma_T$$

3. Comme σ_R appartient aux deux ensembles, il suffit de montrer la transitivité et l'augmentation à l'aide de la réflexivité et la pseudo-transitivité seulement. La transitivité est en fait un cas dégénéré de la pseudo-transitivité avec $W = \emptyset$:

$$\dfrac{X \to Y \qquad Y \to Z}{X \to Z}\,\sigma_P$$

   Pour l'augmentation s'obtient en posant $Z = WY$ dans la règle de pseudo-transitivité :

$$\dfrac{X \to Y \qquad \dfrac{WY \subseteq WY}{WY \to WY}\,\sigma_R}{WX \to WY}\,\sigma_P$$

4. L'antépénultième question montre que le système $\{\sigma_R, \sigma_P\}$ est correct. D'autre part, on sait que le système d'Armstrong $\{\sigma_R, \sigma_A, \sigma_T\}$ est complet, la question précédente montre que on peut transformer toutes les preuves de ce système par des preuves ne faisant intervenir que $\{\sigma_R, \sigma_P\}$ ce qui montre sa complétude.

Solution de l'exercice 3

1. **Extensive** Soit $A \in X$, par l'axiome de réflexivité on a $\vdash X \to A$ et donc $A \in \{A \mid \Sigma \vdash X \to A\}$.

   Croissante Soit $X \subseteq Y$, il faut montrer que $X^\star \subseteq Y^\star$. Considérons $A \in X^\star$, par définition, il existe une preuve de $\Sigma \vdash X \to A$. Comme $X \subseteq Y$, par l'axiome de réflexivité on a $\vdash Y \to X$. Par transitivité on a une preuve de $\Sigma \vdash Y \to A$, c'est-à-dire $A \in Y^\star$.

   Idempotente On montre la double inclusion. Pour la première direction, on a $X \subseteq X^\star$ et par monotonie $X^\star \subseteq (X^\star)^\star$. Pour la seconde direction, il faut prouver que $(X^\star)^\star \subseteq X^\star$. On prouve d'abord que $\Sigma \vdash X \to X^\star$. Supposons sans perte de généralités que $X^\star = A_1 \dots A_n$, pour tout indice $1 \leq i \leq n$, par définition, si $A_i \in X^\star$ alors on a une preuve que $\Sigma \vdash X \to A_i$. Or on peut concaténer toutes ces preuves puis répéter la règle de composition[2] $\{X \to Y, X \to Z\} \vdash X \to YZ$ pour obtenir une preuve de $\Sigma \vdash X \to X^\star$. Considérons $A \in (X^\star)^\star$, par définition, on a une preuve $\Sigma \vdash X^\star \to A$ (notez le $X^\star$), comme on a prouvé que $\Sigma \vdash X \to X^\star$, on peut conclure par transitivité.

2. Tout d'abord, on sait que $X^\star = X^+$ et on raisonnera sur $X^\star$. D'une part, par hypothèse $X \subseteq Y$ et donc par monotonie on a $X^+ \subseteq Y^+$. D'autre par, par hypothèse, $Y \subseteq X^+$ et donc par monotonie puis par idempotence on a $Y^+ \subseteq (X^+)^+ = X^+$. On a donc $X^+ \subseteq Y^+ \subseteq X^+$ et on conclut par antisymétrie de la relation d'inclusion.

3. On procède par niveaux : calculer les fermetures de tous les singletons, puis des couples d'attributs, puis des ensembles de trois attributs etc. Au total, il y a $2^5 = 32$ sous-ensembles dont il faudrait calculer la fermeture. On rappelle que le nombre de combinaisons de k parmi n est $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

niveau 1	
$A^+ = A$	
$B^+ = BDE$	
$C^+ = CE$	
$D^+ = BDE$	
$E^+ = E$	
niveau 2	
$AB^+ = R$	
$AC^+ = R$	
$AD^+ = R$	
$AE^+ = R$	
$BC^+ = R$	
$BD^+ = BDE$	(1)
$BE^+ = BDE$	(2)
$CD^+ = R$	
$CE^+ = CE$	(3)

   Pour (1) et (2), comme $B \subseteq BD \subseteq B^+ = BDE$, on déduit que $BD^+ = BDE$. De la même façon pour (3), en remarquant que $C \subseteq CE \subseteq C^+ = CE$, on déduit $CE^+ = CE$. Pour le niveau 3, on doit énumérer les $\binom{5}{2} = 10$ ensembles de taille 3, mais en remarquant que $AB^+ = AC^+ = AD^+ = AE^+$ on peut éliminer tous ceux qui contiennent A. Il en reste donc $\binom{4}{3} = 4$, à savoir $\{BCD, BDE, BCE, CDE\}$

niveau 3	
$BCD^+ = R$	(4)
$BDE^+ = BDE$	(5)
$BCE^+ = R$	(6)
$CDE^+ = R$	(7)

   Pour (4) et (7), on remarque que BCD^+ et CDE^+ contiennent CD et donc leur fermeture est R. Pour (5), on le connait déjà, et enfin pour (6), on utilise le fait que $BC^+ = R$. On a terminé, car il ne reste que $BCDE$ à tester (le seul sous-ensmeble de taille 4 qui ne contienne pas A) et on sait que $BCDE^+ = R$. En conclusion, on a :

$$Cl(\Sigma) = \{A, E, CE, BDE, R\}$$

[2] cette dernière étant simplement une combinaison d'augmentation et de transitivité.

Solution de l'exercice 4

1. Le sens (1) $\Rightarrow$ (2) est clair car si $\Sigma \models \alpha$ pour toute relation r, alors c'est aussi vrai pour le cas limité à $|r| = 2$. Pour la réciproque (2) $\Rightarrow$ (1), supposons le contraire, c'est-à-dire $\Sigma \models_2 \alpha$ et $\Sigma \not\models \alpha$, pour aboutir à une contradiction. Si $\Sigma \not\models \alpha$, alors il existe une relation r (de taille arbitraire) telle que $r \models \Sigma$ et $r \not\models \alpha$. Cela signifie qu'ils existent (au moins) deux tuples $t_1, t_2 \in r$ tels que $t_1[A_1 \dots A_p] = t_2[A_1 \dots A_p]$ et $t_1[B_1 \dots B_q] \neq t_2[B_1 \dots B_q]$. Il suffit donc de considérer la relation $r' = \{t_1, t_2\}$ qui a exactement deux tuples : elle vérifie Σ par hypothèse mais on vient de montrer qu'elle viole α, contradiction avec $\Sigma \models_2 \alpha$.
2. Pour la direction *si*. Supposons $r \models \alpha$, deux cas sont possibles :
    1. α est trivialement satisfaite car $t_1[A_i] \neq t_2[A_i]$ pour un certain i. Dès lors $\nu \not\models \underline{A}_1 \wedge \dots \wedge \underline{A}_p$ car $\nu(\underline{A}_i) = 0$ et ainsi $\nu \models \underline{A}_1 \wedge \dots \wedge \underline{A}_p \Rightarrow \underline{B}_1 \wedge \dots \wedge \underline{B}_q$ par inspection de la table de vérité de $\Rightarrow$.
    2. α est satisfaite car $t_1[A_i] = t_2[A_i]$ pour tout i et de même $t_1[B_j] = t_2[B_j]$. On a donc $\nu \models \underline{A}_1 \wedge \dots \wedge \underline{A}_p$ et également $\nu \models \underline{B}_1 \wedge \dots \wedge \underline{B}_q$. On conclut encore par inspection de la table de vérité de $\Rightarrow$.

    Pour la direction *seulement si*, on suppose que $\nu \models_P \underline{\alpha}$ et on considère encore les deux cas :

    1. soit $\nu \models_P \underline{\alpha}$ par vacuité car $\nu(\underline{A}_i) = 0$ pour un certain i et donc dans ce cas α est trivialement satisfaite car $t_1[A_i] \neq t_2[A_i]$,
    2. soit on a $\nu(\underline{A}_i) = 1$ pour tout i et aussi $\nu(\underline{B}_j) = 1$ pour chaque j car $\nu \models_P \underline{\alpha}$. Dans ce cas aussi α est satisfaite car $\{t_1, t_2\}$ est le seul couple de tuple à considérer pour vérifier $r \models \alpha$.
3. Pour le sens (3) $\Rightarrow$ (2) on procède par l'absurde. On suppose donc $\Sigma \models_2 \alpha$ et $\underline{\Sigma} \not\models_P \underline{\alpha}$. Il existe donc une valuation ν telle que $\nu \models_P \underline{\Sigma}$ et $\nu \not\models_P \underline{\alpha}$. Par inspection de la table de vérité de $\Rightarrow$, cela signifie que $\nu(\underline{A}_i) = 1$ pour tout i mais qu'il existe (au moins) un index j tel que $\nu(\underline{B}_j) = 0$. A partir de cette valuation, construisons alors une relation $r = \{t_1, t_2\}$ telle que décrit dans le lemme 1. D'après le lemme on a $r \not\models \alpha$ et $r \models \Sigma$, contradiction.
   Pour le sens (2) $\Rightarrow$ (3), on a procéder de façon similaire en supposant $\underline{\Sigma} \models_P \underline{\alpha}$ et $\Sigma \not\models_2 \alpha$. On a donc au moins une relation à deux tuples $r = \{t_1, t_2\}$ qui vérifie Σ mais pas α. On construit alors la valuation ν telle que $\nu(\underline{A}) = 1$ si et seulement si $t_1[A] = t_2[A]$. On est dans le cas d'application du lemme 1 et d'après celui-ci on déduit que $\nu \not\models \underline{\alpha}$ et $\nu \models \underline{\Sigma}$, une contradiction.
4. D'après la question 1 on a (1) $\equiv$ (3), d'après la question 3 on a aussi (2) $\equiv$ (3), les trois propositions sont donc équivalentes.