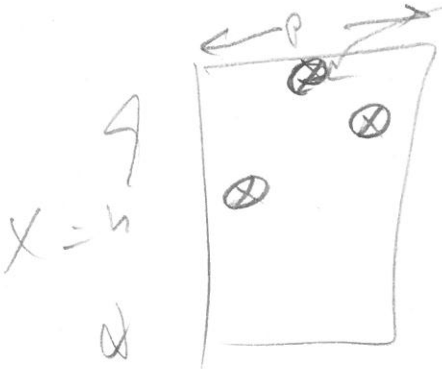
Math 340.4 Lec 25 5/10/12

Missingness - scourge of data analysis



for obs. unit 1, the 9th feature's value is not present.

You cannot use any of the models in this class with this. Let M_j denote Bern. r.v. that jth feature is missing

Types of missingness: missing data mechanism, MDM

MDM	$P(M_j \mid X_{j,miss}, X_{-j,miss}, X_{-j,obs}, V, \gamma)$ = . . (V: other obs. vars, γ: other params)
MCAR	$P(M_j \mid \gamma)$
MAR	$P(M_j \mid X_{-j,miss}, X_{j,obs})$
NMAR	does not simplify

examples of each?

MCAR: data corruption

MAR: old person surveys

NMAR: chicago school showing GPA, age $X = \{GPA, age\}$

How to handle?

(I) Listwise deletion: drop all units. Bad for
(a) pred. performance
(b) generalization ... my predictions will be overconfident

(II) Imputation: fill in holes with values. Strategies
(a) use $\bar{X}_j$ for any value X_{ij} missing. Not great!
(b) fit a model to X_j as outcome and predictors, make better!
"X_{imp}"

Rec: MissForest: alg. which iteratively runs RF on all features filling in holes until "convergence"!

$Y = f(X_1, \dots X_p) + \varepsilon$ but it also may be ← bring units for missingness

$Y = f(X_1, \dots, X_p, m_1, \dots, m_p) + \varepsilon$ value missingness itself matters.

e.g. Chicago public school.

Rec: create dummies $m_1, \dots m_p$, then impute then your full design matrix is

$$X_{full} = \underbrace{[X_{imp}, m_1, m_2, \dots, m_p]}_{2p \text{ terms}}$$

Question if X_j is normal, you could just add a level for m_j and not impute...

Credit Card Example

How much will individual pay back?
Will the individual fully pay back or not?
What is the prob (↗) ?

Type of Supervised Learning

Regression
Binary Classification
Prob. estimation / prob. class.

previously if $y = \{0,1\}$

$$y = t(z_1, \dots z_t) = f(x_1, \dots x_p) + \varepsilon = h^*(x_1, \dots x_p) + \varepsilon = g(x_1, \dots x_p) + e$$

$\in \{0,1\}$ $\in \{0,1\}$ $\in \{-1, 0, 1\}$ ditto

t, f, h^*, g are all outputting 0,1 ; not probs!

Consider an alternate construction of f, h^*, g but not t! Let $Y_1, \dots, Y_n$ be ind. Bernoullis with diff θ's

$Y_i \sim \text{Bernoulli}\left(\underset{pr}{f}(x_1, \dots x_p)\right)$ ← $\text{Prob}(Y_i = 1 \mid \vec{X})$ the best you can do with the data at hand

$Y_i \sim \text{Bernoulli}\left(h^*_{pr}(x_1, \dots x_p)\right)$ is a worse model

$Y_i \sim \text{Bernoulli}\left(g_{pr}(x_1, \dots x_p)\right)$ is even worse

Is there a t_{pr}? Yes $t_{pr} = t$ since if all covariates known,

$\Rightarrow$ no randomness and $\underset{\in\{0,1\}}{\text{prob}(+ \mid \vec{x})} = \underset{\in\{0,1\}}{y}$ itself

$t - f_{pr}$: error due to ignorance
$f_{pr} - h^*_{pr}$: error due to noise
$h^*_{pr} - g_{pr}$: " " regression error

prob = 100% $\Rightarrow$ prob = 90%

What is $g_{pr,0}$? The null model? $\bar{y} = \frac{\Sigma y}{n} = \hat{p}$

As the model gets worse... it runs towards $g_{pr,0}$.

How to fit and get g_{pr}?

Assume that each $(\vec{x}_i, y) \in D$ are independent.

then $P(D) = P((\vec{x}_1, y_1)) P((\vec{x}_2, y_2)) \cdots P((\vec{x}_n, y_n))$ (from Prob 281...)

$= g_{pr}(\vec{x}_1)^{y_1}(1 - g_{pr}(\vec{x}_1))^{1-y_1} (\quad)(\quad)\cdots(\quad)$

$= \prod_{i=1}^{n} g_{pr}(\vec{x}_i)^{y_i}\left(1 - g_{pr}(\vec{x}_i)\right)^{1-y_i}$

Now we want to find g_{pr} s.t. $P(D)$ is maximal.

Of course g_{pr} is very complicated. So... we make as assumption of $\mathcal{H}_{pr}$. What is an example? We need functions $h_{pr}: \mathbb{R}^p \to$? $[0,1]$. How about $\{\Pi_{\vec{w}} : \vec{w} \cdot \vec{x} > 0, \vec{w} \in \mathbb{R}^{p+1}\}$

Not good! Only returns $\{0,1\}$. Prob of 50%, 59%, 93% not possible.

How about $\mathcal{H} = \{\vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{p+1}\}$? $\vec{w} \cdot \vec{x} \in \mathbb{R}$. We can return probs of 107%! or -38%...

What if we wanted to return the linear model but output in $\in (0,1)$?

Every single link function: $\mathcal{H} = \{\phi(\vec{w}\cdot\vec{x}) : \vec{w} \in \mathbb{R}^{p+1}\}$

$\phi: \mathbb{R} \to (0,1)$ and monotonic and $\phi' > 0$ always!

if $\vec{w}\cdot\vec{x} \uparrow \Rightarrow P(Y=1|\vec{x}) \uparrow$ Also $P(Y=1|\vec{x}) \neq 1$, $\neq 0$

why? Can never be completely sure!

The link function we will study is the **logistic link**:

$$\phi(\eta) = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}}$$ using this is logistic regression

Fact: Another one is **probit**

$$\phi(\eta) = \underbrace{F^{-1}(\eta)}_{\text{inverse CDF of standard normal}}$$

Fact: or **complementary log-log**:

$$\phi(\eta) = 1 - e^{-e^\eta}$$

$$\mathcal{H} = \left\{ \frac{e^{\vec{w}\cdot\vec{x}}}{1+e^{\vec{w}\cdot\vec{x}}} : \vec{w} \in \mathbb{R}^{p+1} \right\}$$

$\mathcal{L}$

$$\Rightarrow \vec{b} = \underset{\vec{w}\in\mathbb{R}^{p+1}}{\operatorname{argmax}} \left\{ \prod_{i=1}^{n} \left(\frac{e^{\vec{w}\cdot\vec{x}_i}}{1+e^{\vec{w}\cdot\vec{x}_i}}\right)^{y_i} \left(1 - \frac{e^{\vec{w}\cdot\vec{x}_i}}{1+e^{\vec{w}\cdot\vec{x}_i}}\right)^{1-y_i} \right\}$$

$$\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\vec{w}\cdot\vec{x}_i}}\right)^{y_i} \left(\frac{1}{1+e^{\vec{w}\cdot\vec{x}}}\right)^{1-y_i}$$

We cannot take $\frac{d\mathcal{L}}{d\vec{w}} \stackrel{\text{set}}{=} 0$ and solve for $\vec{w}$ we need to use opt. alg. e.g. grad. descent, just like SVM.

Using numerical methods, we get $\vec{b}$.

$$\Rightarrow g_{pr}(\vec{x}) = \frac{e^{\vec{b}\cdot\vec{x}}}{1+e^{\vec{b}\cdot\vec{x}}} = \text{prob est's } \hat{P}(Y=1 \mid \vec{x})$$

What does $\vec{b}\cdot\vec{x}$ mean?

Let's see...

$$\hat{p} = \phi(\vec{b}\cdot\vec{x}) = \frac{1}{1+e^{-\vec{b}\cdot\vec{x}}}$$

$$\Rightarrow \frac{1}{\hat{p}} = 1+e^{-\vec{b}\cdot\vec{x}} \Rightarrow \frac{1}{\hat{p}} - 1 = e^{-\vec{b}\cdot\vec{x}}$$

$\swarrow$ Odds$(Y=1 \mid \vec{x})$

$$\Rightarrow \frac{1-\hat{p}}{\hat{p}} = e^{-\vec{b}\cdot\vec{x}} \Rightarrow \frac{\hat{p}}{1-\hat{p}} = e^{\vec{b}\cdot\vec{x}} \Rightarrow \vec{b}\cdot\vec{x} = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$$

$$= \log(\text{odds}(Y=1 \mid \vec{x}))$$

Interpret b_j! Same as last class except change "y increase by b_j" to "log odds of $P(Y=1)$ increase by b_j".

Validating g_{pr}. What is best prob fcn? f_{pr}. So let's validate against this! You can't! Unknown!

The only thing you have is $y_1, \dots, y_n \in \mathcal{Y} = \{0, 1\}$