LEC003 Demand Forecasting

VG441 SS2021

Cong Shi
Industrial & Operations Engineering
University of Michigan

ARMA(p,q)

- AR(p)

$$X_t = \mu + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

- MA(q)

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

- ARMA(p,q)

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

ARIMA(p,d,q)

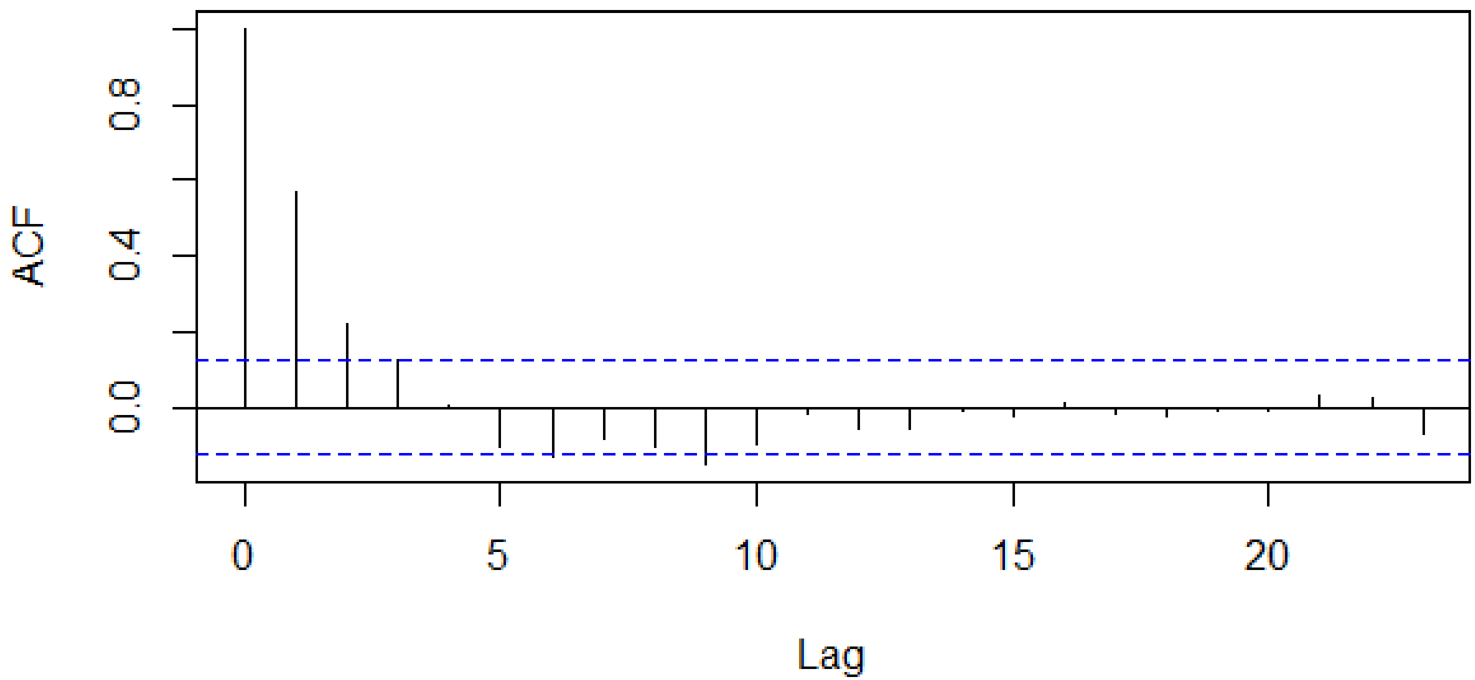
- d is the degree of difference, e.g., d =1

$$Y_t = X_t - X_{t-1}$$

$$Y_t = \mu + \varepsilon_t + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

- How about d = 2?

ACF/PACF



Python Time!

- statsmodels.tsa.arima_model

Logistic Regression (0-1)

- Given features *x*, predict either 1 or 0 (on or off)

p

sigmoid/logistic function

$$p(1 \mid \boldsymbol{x}, \boldsymbol{w}) := \sigma(\boldsymbol{w} \cdot \boldsymbol{x}) := \frac{1}{1 + \exp(-\boldsymbol{w} \cdot \boldsymbol{x})}$$

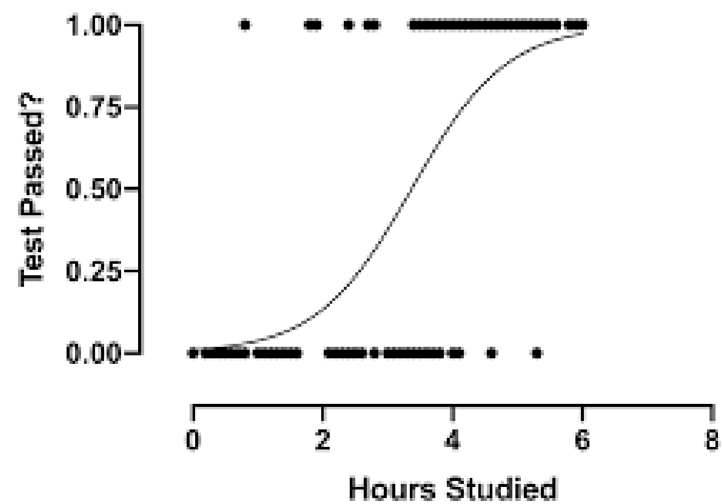
1-p

$$p(0 \mid \boldsymbol{x}, \boldsymbol{w}) = 1 - \sigma(\boldsymbol{w} \cdot \boldsymbol{x}) = \frac{\exp(-\boldsymbol{w})}{1 + \exp(-\boldsymbol{w} \cdot \boldsymbol{x})}$$

Connection to linear regression…

$$\log\left(\frac{p}{1-p}\right) = \boldsymbol{w} \cdot \boldsymbol{x}$$

logit function/log odds

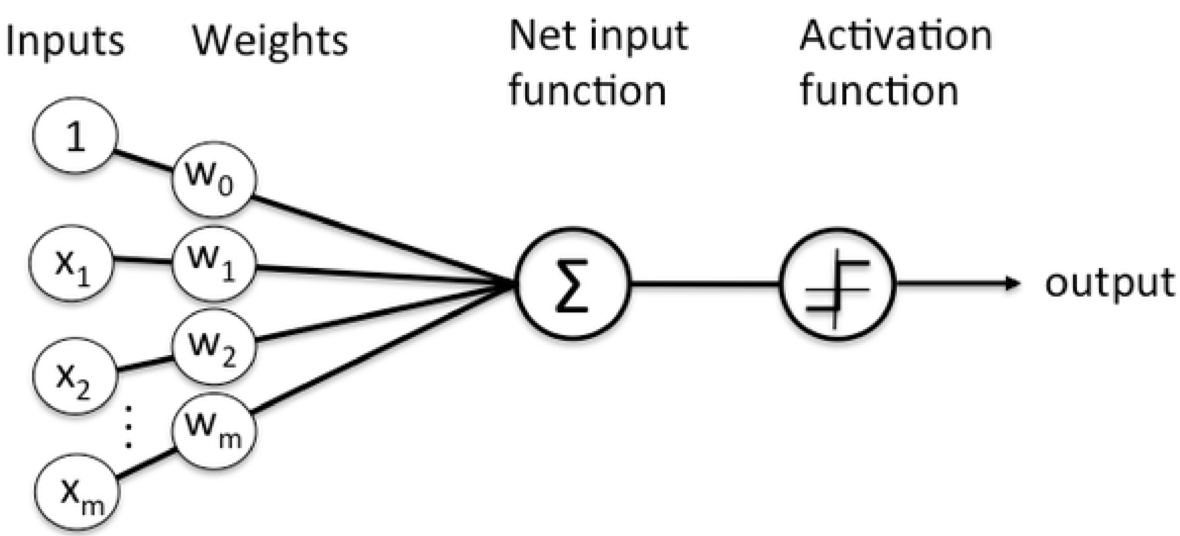


Logistic Regression (1-layer NN)

- Given features *x*, predict either 1 or 0 (on or off)

$\boxed{p}$ $\boxed{\text{sigmoid/logistic function}}$

$$p(1 \mid \boldsymbol{x}, \boldsymbol{w}) := \sigma(\boldsymbol{w} \cdot \boldsymbol{x}) := \frac{1}{1 + \exp(-\boldsymbol{w} \cdot \boldsymbol{x})}$$



Logistic Regression (0-1)

- Given features *x*, predict either 1 or 0 (on or off)

$$p(1 \mid \boldsymbol{x}, \boldsymbol{w}) := \sigma(\boldsymbol{w} \cdot \boldsymbol{x}) := \frac{1}{1 + \exp(-\boldsymbol{w} \cdot \boldsymbol{x})}$$

- Minimizing cross-entropy loss function:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} \left(-y^{(i)} \log \sigma(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)}) - (1 - y^{(i)}) \log \sigma(-\boldsymbol{w} \cdot \boldsymbol{x}^{(i)}) \right)$$

Probabilistic Interpretation

- Given features *x*, predict either 1 or 0 (on or off)

$$p(1 \mid \boldsymbol{x}, \boldsymbol{w}) := \sigma(\boldsymbol{w} \cdot \boldsymbol{x}) := \frac{1}{1 + \exp(-\boldsymbol{w} \cdot \boldsymbol{x})}$$

- Minimizing the "negative" MLE:

$$J_S^{\mathrm{LOG}}(\boldsymbol{w}) := -\frac{1}{m} \sum_{i=1}^{m} \log p\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{w}\right)$$

$$\boxed{\begin{aligned} -\log p(y \mid \boldsymbol{x}, \boldsymbol{w}) &= -y \log \sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - (1-y) \log \sigma(-\boldsymbol{w} \cdot \boldsymbol{x}) \\ &= \begin{cases} \log(1 + \exp(-\boldsymbol{w} \cdot \boldsymbol{x})) & \text{if } y = 1 \\ \log(1 + \exp(\boldsymbol{w} \cdot \boldsymbol{x})) & \text{if } y = 0 \end{cases} \end{aligned}}$$

$$\nabla_{\boldsymbol{w}}(-\log p(y \mid \boldsymbol{x}, \boldsymbol{w})) = -(y - \sigma(\boldsymbol{w} \cdot \boldsymbol{x}))\boldsymbol{x}$$

Gradient Descent

- Minimizing the "negative" MLE:

$$J_S^{\mathrm{LOG}}(\boldsymbol{w}) := -\frac{1}{m}\sum_{i=1}^{m} \log p\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{w}\right)$$

- Gradient?

$$\nabla J_S^{\mathrm{LOG}}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - \sigma\left(\boldsymbol{w}\cdot\boldsymbol{x}^{(i)}\right)\right)\boldsymbol{x}^{(i)}$$

$$\boxed{\begin{aligned} -\log p(y \mid \boldsymbol{x}, \boldsymbol{w}) &= -y\log\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - (1-y)\log\sigma(-\boldsymbol{w}\cdot\boldsymbol{x}) \\ &= \begin{cases} \log(1+\exp(-\boldsymbol{w}\cdot\boldsymbol{x})) & \text{if } y = 1 \\ \log(1+\exp(\boldsymbol{w}\cdot\boldsymbol{x})) & \text{if } y = 0 \end{cases} \end{aligned}}$$

$$\nabla_{\boldsymbol{w}}(-\log p(y \mid \boldsymbol{x}, \boldsymbol{w})) = -(y - \sigma(\boldsymbol{w}\cdot\boldsymbol{x}))\boldsymbol{x}$$

Gradient Descent

Input: training objective

$$J_S^{\mathrm{LOG}}(\boldsymbol{w}) := -\frac{1}{m}\sum_{i=1}^{m}\log p\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{w}\right)$$

Output: parameter $\hat{\boldsymbol{w}} \in \mathbb{R}^n$ such that $J_S^{\mathrm{LOG}}(\hat{\boldsymbol{w}}) \approx J_S^{\mathrm{LOG}}\left(\boldsymbol{w}_S^{\mathrm{LOG}}\right)$

1. Initialize $\boldsymbol{w}^0$ (e.g., randomly).
2. For $t = 0 \ldots T-1$,

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + \frac{\eta^t}{m}\sum_{i=1}^{m}\left(y^{(i)} - \sigma\left(\boldsymbol{w}^t \cdot \boldsymbol{x}^{(i)}\right)\right)\boldsymbol{x}^{(i)}$$

3. Return $\boldsymbol{w}^T$.

From Binary to Multinomial

- Given features x, predict $y \in \{1, \ldots, k\}$

weights for each class

P(y)

$$p(y \mid \boldsymbol{x}, \theta) = \frac{\exp\left(\boldsymbol{w}^{y} \cdot \boldsymbol{x}\right)}{\sum_{y'=1}^{k} \exp\left(\boldsymbol{w}^{y'} \cdot \boldsymbol{x}\right)}$$

softmax function

From Binary to Multinomial

- Minimize cross-entropy loss function

$$J_S^{\mathrm{LLM}}(\boldsymbol{w}) := -\frac{1}{m}\sum_{i=1}^{m}\log p\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \theta\right)$$

$$\boxed{-\log p(y \mid \boldsymbol{x}, \theta) = \underbrace{\log\left(\sum_{y'=1}^{k}\exp\left(\boldsymbol{w}^{y'}\cdot\boldsymbol{x}\right)\right)}_{\text{constant wrt. } y} - \underbrace{\boldsymbol{w}^{y}\cdot\boldsymbol{x}}_{\text{linear}}}$$

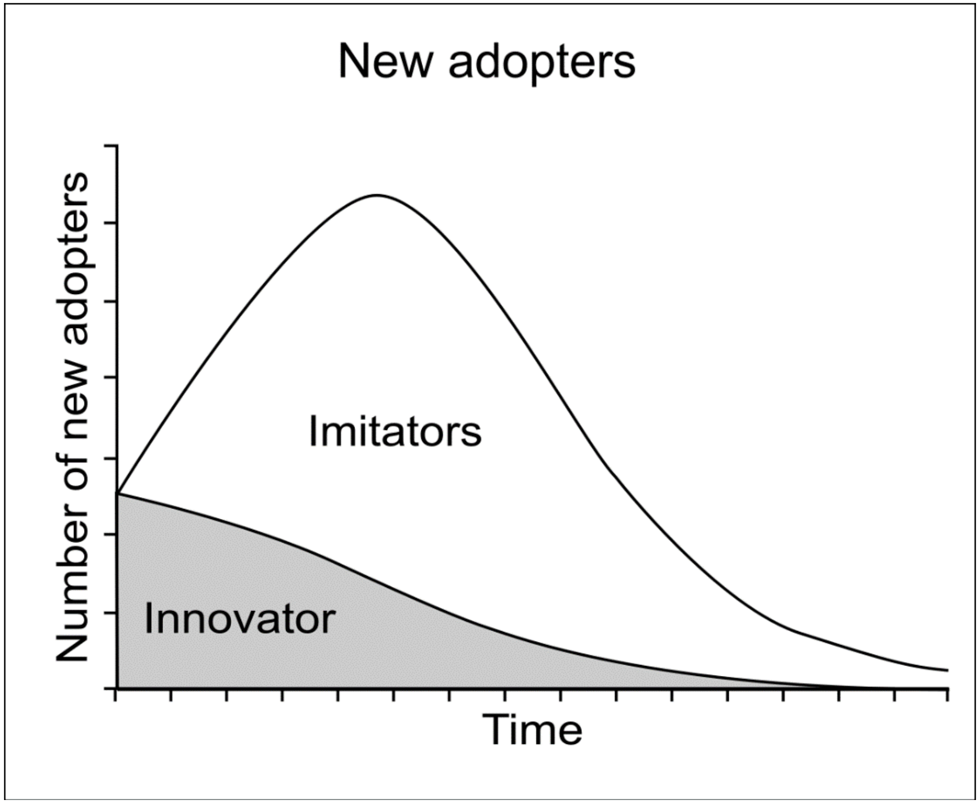
$$\nabla_{\boldsymbol{w}^l}(-\log p(y \mid \boldsymbol{x}, \theta)) = \begin{cases} -(1-p(l \mid \boldsymbol{x}, \theta))\boldsymbol{x} & \text{if } l = y \\ p(l \mid \boldsymbol{x}, \theta)\boldsymbol{x} & \text{if } l \neq y \end{cases}$$

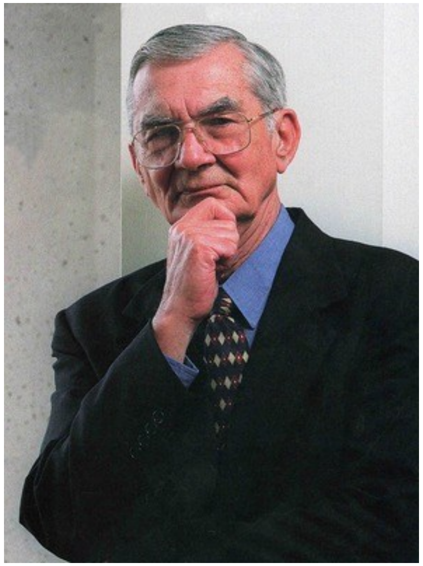
Python Time!

- from sklearn.linear_model import LogisticRegression

Bass Diffusion Model

- How about projecting sales of new products?





Frank Bass

Bass Diffusion Model

Cumulative purchase probability of a random customer $F(t)$

Purchase probability at time t $f(t) = F'(t)$

The rate of purchase at time t (given no purchase so far)

$$\frac{f(t)}{1 - F(t)} = \underbrace{p}_{\text{Coefficient of innovation}} + \underbrace{q}_{\text{Coefficient of imitation}} F(t)$$

Bass Solution

$$\frac{dF/dt}{1-F} = p + qF$$

$$\frac{dF}{dt} = p + (q-p)F - qF^2$$

$$\int \frac{1}{p + (q-p)F - qF^2} dF = \int dt$$

$$\begin{aligned}
\frac{1}{(p+qF)(1-F)} &= \frac{A}{p+qF} + \frac{B}{1-F} \\
&= \frac{A - AF + pB + qFB}{(p+qF)(1-F)} \\
&= \frac{A + pB + F(qB - A)}{(p+qF)(1-F)}
\end{aligned}$$

$$A = q/(p+q)$$
$$B = 1/(p+q)$$

Bass Solution

$$\int \frac{1}{(p+qF)(1-F)} dF = \int dt$$

$$\int \left(\frac{A}{p+qF} + \frac{B}{1-F} \right) dF = t + c_1$$

$$\int \left(\frac{q/(p+q)}{p+qF} + \frac{1/(p+q)}{1-F} \right) dF = t + c_1$$

$$\frac{1}{p+q} \ln(p+qF) - \frac{1}{p+q} \ln(1-F) = t + c_1$$

$$\frac{\ln(p+qF) - \ln(1-F)}{p+q} = t + c_1$$

Boundary Condition

$$t = 0 \quad \Rightarrow \quad F(0) = 0$$

$$t = 0 \quad \Rightarrow \quad c_1 = \frac{\ln p}{p+q}$$

$$F(t) = \frac{p\left(e^{(p+q)t} - 1\right)}{pe^{(p+q)t} + q}$$

Bass Solution

Calibration

- Sales in any period are $s(t) = mf(t)$
- Cumulative sales up to time t are $S(t) = mF(t)$

$$\frac{s(t)/m}{1 - S(t)/m} = p + qS(t)/m$$

$$s(t) = [p + qS(t)/m][m - S(t)]$$

$$s(t) = \beta_0 + \beta_1 S(t) + \beta_2 S(t)^2 \quad (BASS)$$

$$\beta_0 = pm$$

$$\beta_1 = q - p$$

$$\beta_2 = -q/m$$

$$\Rightarrow$$

$$m = \frac{-\beta_1 \pm \sqrt{\beta_1^2 - 4\beta_0\beta_2}}{2\beta_1}$$

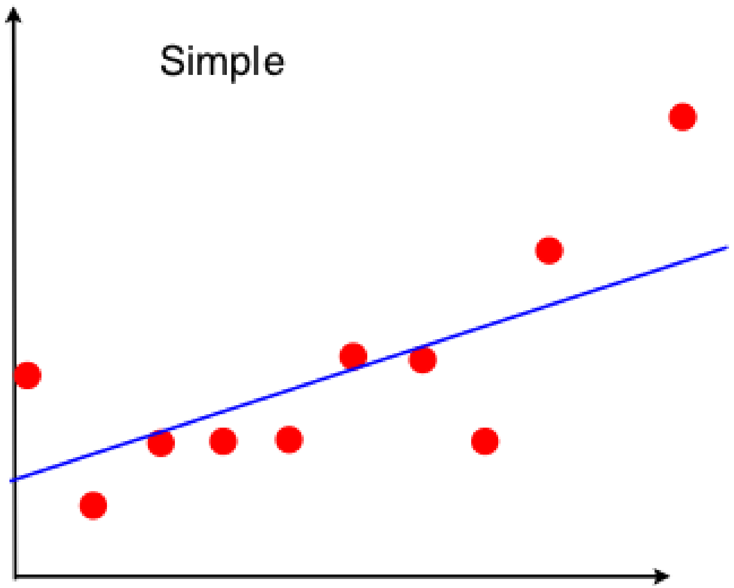
$$p = \frac{\beta_0}{m}; \quad q = -m\beta_2$$

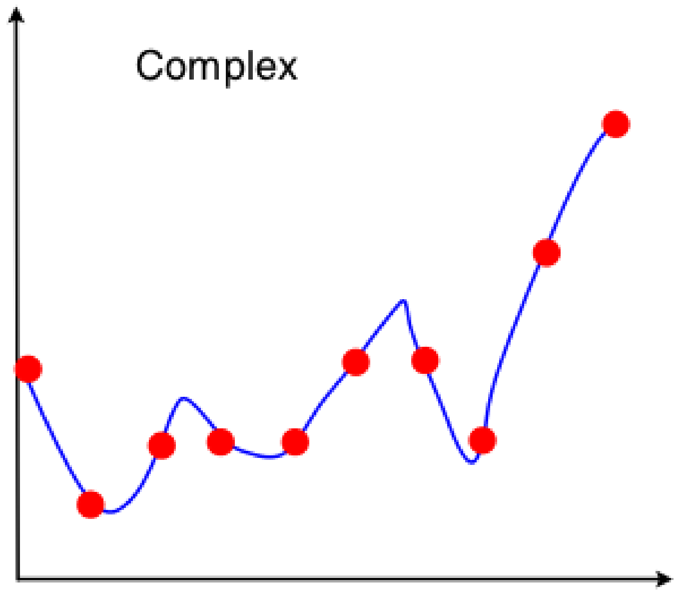
Conduct a linear regression!

Python Time!

- from sklearn import linear_model
- from statsmodels.api import OLS

Bias-Variance Tradeoff





Kaggle Competition

All Competitions

Active Completed InClass

All Categories Default Sort

Jigsaw Multilingual Toxic Comment Classification Use TPUs to identify toxicity comments across multiple languages Featured • a month to go • Code Competition • 862 Teams	$50,000
M5 Forecasting - Accuracy Estimate the unit sales of Walmart retail goods Featured • 2 months to go • 3589 Teams	$50,000
M5 Forecasting - Uncertainty Estimate the uncertainty distribution of Walmart unit sales. Featured • 2 months to go • 389 Teams	$50,000
University of Liverpool - Ion Switching Identify the number of channels open at each time point Research • 16 days to go • 2333 Teams	$25,000
TReNDS Neuroimaging Multiscanner normative age and assessments prediction with brain function, structure, and connectivity Research • 2 months to go • 275 Teams	$25,000
ALASKA2 Image Steganalysis Detect secret data hidden within digital images Research • 2 months to go • 237 Teams	$25,000
Prostate cANcer graDe Assessment (PANDA) Challenge Prostate cancer diagnosis using the Gleason grading system Featured • 2 months to go • Code Competition • 309 Teams	$25,000