More Submodular stuff

D.A. Forsyth, working entirely from Carlos Guestrin's slides

sense
learn
act

Example: Submodularity of info-gain

$Y_1,\ldots,Y_m$, X_1, …, X_n discrete RVs

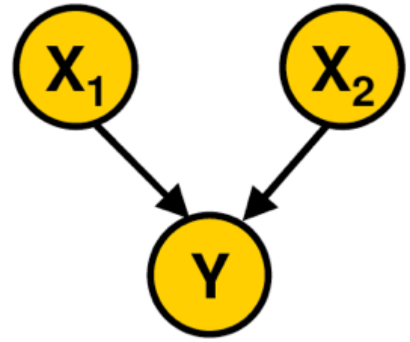
$F(A) = IG(Y; X_A) = H(Y)-H(Y \mid X_A)$

- F(A) is always monotonic
- However, NOT always submodular

An “elementary” counterexample

$X_1, X_2 \sim$ Bernoulli(0.5)

$Y = X_1$ **XOR** X_2



Let $F(A) = IG(X_A; Y) = H(Y) - H(Y|X_A)$

$Y \mid X_1$ and $Y \mid X_2$ $\sim$ Bernoulli(0.5) (entropy 1)

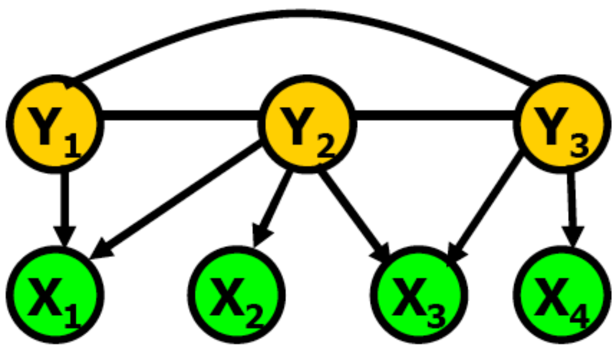
$Y \mid X_1, X_2$ is deterministic! (entropy 0)

Hence $F(\{1,2\}) - F(\{1\}) = 1$, but

$F(\{2\}) - F(\emptyset) = 0$

F(A) submodular under some conditions! (later)

Theorem [Krause & Guestrin UAI' 05]
If X_i are all conditionally independent given Y, then F(A) is submodular!



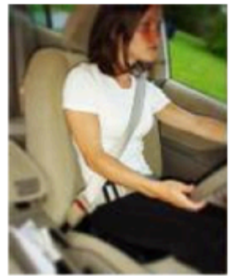
Hence, greedy algorithm works!

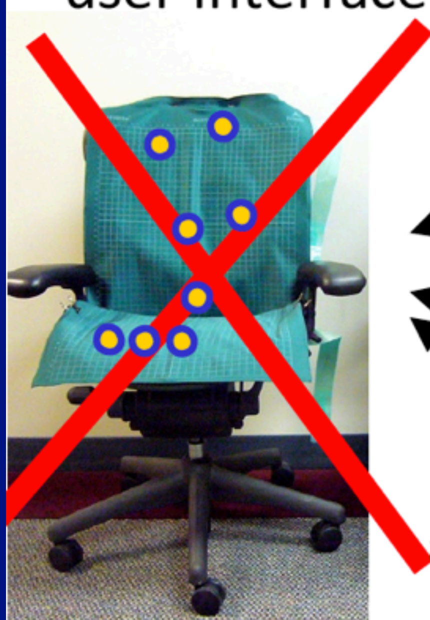
In fact, NO algorithm can do better than (1-1/e) approximation!

Building a Sensing Chair

[Mutlu, Krause, Forlizzi, Guestrin, Hodgins UIST '07]

- People sit a lot
- Activity recognition in assistive technologies
- Seating pressure as user interface

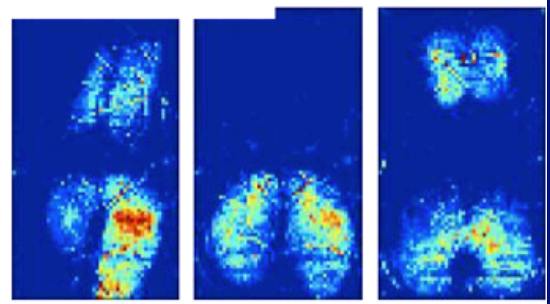




Equipped with 1 sensor per cm^2!

Costs $16,000! ☹

Can we get similar accuracy with fewer, cheaper sensors?



Lean left | Lean forward | Slouch

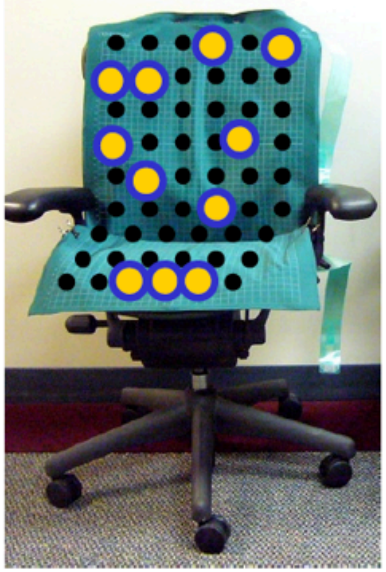
82% accuracy on 10 postures! [Tan et al]

How to place sensors on a chair?

- Sensor readings at locations V as random variables
- Predict posture Y using probabilistic model P(Y,V)
- Pick sensor locations A* $\subseteq$ V to minimize entropy:

$$\boxed{\mathcal{A}^* = \operatorname*{argmax}_{|\mathcal{A}|\leq k} \mathrm{IG}(Y; X_{\mathcal{A}})}$$

Possible locations **V**



Placed sensors, did a user study:

	Accuracy	Cost
Before	82%	$16,000 ☹
After		

Similar accuracy at <1% of the cost!

Variance reduction

(a.k.a. Orthogonal matching pursuit, Forward Regression)

- Let $Y = \sum_i \alpha_i X_i + \varepsilon$, and $(X_1,\ldots,X_n,\varepsilon) \sim N(\cdot; \mu,\Sigma)$
- Want to pick subset X_A to predict Y
- $\mathrm{Var}(Y \mid X_A=x_A)$: conditional variance of Y given $X_A = x_A$
- Expected variance: $\mathrm{Var}(Y \mid X_A) = \int p(x_A)\, \mathrm{Var}(Y \mid X_A=x_A)\, dx_A$
- Variance reduction: $F_V(A) = \mathrm{Var}(Y) - \mathrm{Var}(Y \mid X_A)$

$F_V(A)$ is always monotonic

Theorem [Das & Kempe, STOC '08]
$F_V(A)$ is submodular*

*under some conditions on Σ

→ **Orthogonal matching pursuit near optimal!**
[see other analyses by Tropp, Donoho et al., and Temlyakov]

Active learning

- Hoi et al, "Batch mode Active Learning...", ICML'08
- Fisher information matrix
  - - Expected value of Hessian of log-likelihood
  - Big -> log-likelihood is tightly peaked
- Natural criterion for selecting examples to be labelled
  - alpha - classifier parameters
  - p - distribution of labelled examples
  - q - distribution of unlabelled that are chosen for labelling

$$q^* = \arg\min_q \operatorname{tr}(I_q(\alpha)^{-1} I_p(\alpha))$$

Active learning

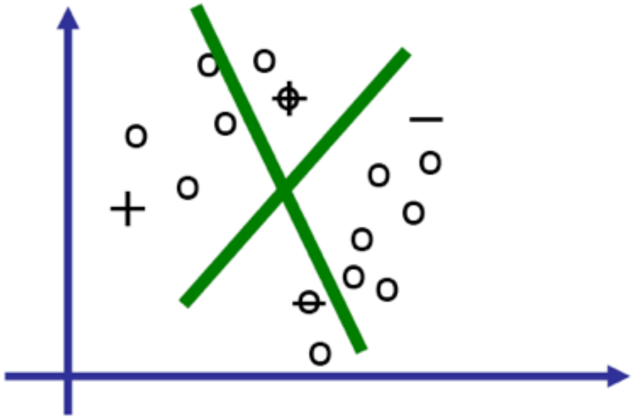
- By a series of approximations, we get

$$\min_{|S|=k \wedge S \subseteq D} \sum_{\mathbf{x} \notin S} \frac{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}{\delta + \sum_{\mathbf{x}' \in S} \pi(\mathbf{x}')(1-\pi(\mathbf{x}'))(\mathbf{x}^\top \mathbf{x}')^2}$$

- Substitute with

$$\begin{aligned} f(S) \quad = \quad & \frac{1}{\delta} \sum_{\mathbf{x} \in D} \pi(\mathbf{x})(1-\pi(\mathbf{x})) \\ & - \sum_{\mathbf{x} \notin S} \frac{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}{\delta + \sum_{\mathbf{x}' \in S} \pi(\mathbf{x}')(1-\pi(\mathbf{x}'))(\mathbf{x}^\top \mathbf{x}')^2} \end{aligned} \tag{6}$$

Batch mode active learning [Hoi et al, ICML'06]



Which data points o should we label to minimize error?

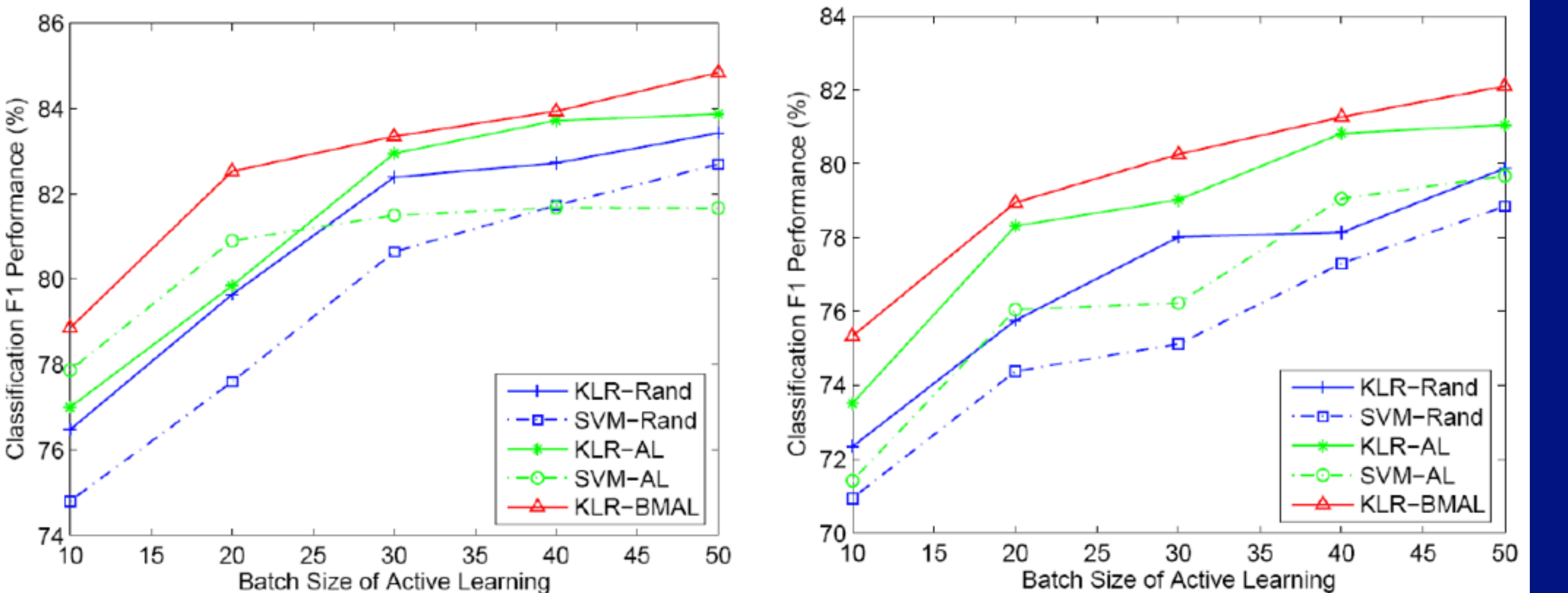
Want batch A of k points to show an expert for labeling

$$F(\mathcal{A}) = \frac{1}{\delta}\sum_{s\in\mathcal{V}}\sigma^2(s) - \sum_{s\notin\mathcal{A}}\frac{\sigma^2(s)}{\delta + \sum_{s'\in\mathcal{A}}\sigma^2(s')(s^T s')}$$

- F(A) selects examples that are
  - uncertain [$\sigma^2(s) = \pi(s)\,(1-\pi(s))$ is large]
  - diverse (points in A are as different as possible)
  - relevant (as close to V\A is possible, s^T s' large)
- F(A) is submodular and monotonic!
  [approximation to improvement in Fisher-information]

Results about Active Learning

[Hoi et al, ICML'06]



(a) Australian (b) Heart

Batch mode Active Learning performs better than

- Picking k points at random
- Picking k points of highest entropy