Pointwise Exact Bootstrap Distributions of Cost Curves

Charles Dugas and David Gadoury

University of Montréal

25th ICML – Helsinki – July 2008

Outline

Introduction

- **Goal**: identifying the presence of a certain condition (e.g. fraud, malignant tumors, defective part, etc.), given a set of features, i.e. binary classification.
- **Model**: outputs a continuous score s for each example of a set. Higher s means higher chances that condition is present.
- **Out-of-sample (OOS) performance**
  - scalars: error rate (accuracy), AUC, etc.
  - curves: ROC, Cost curves
- **Confidence intervals**
  - pointwise: not bands
  - one or two models.

ROC Curves

Threshold t: instance labelled as positive ($s \geq t$) or negative ($s < t$).

Decision	Truth	
	Positive	Negative
Positive	**True positives**	**False positives**
Negative	**False negatives**	**True negatives**

Scalar measures

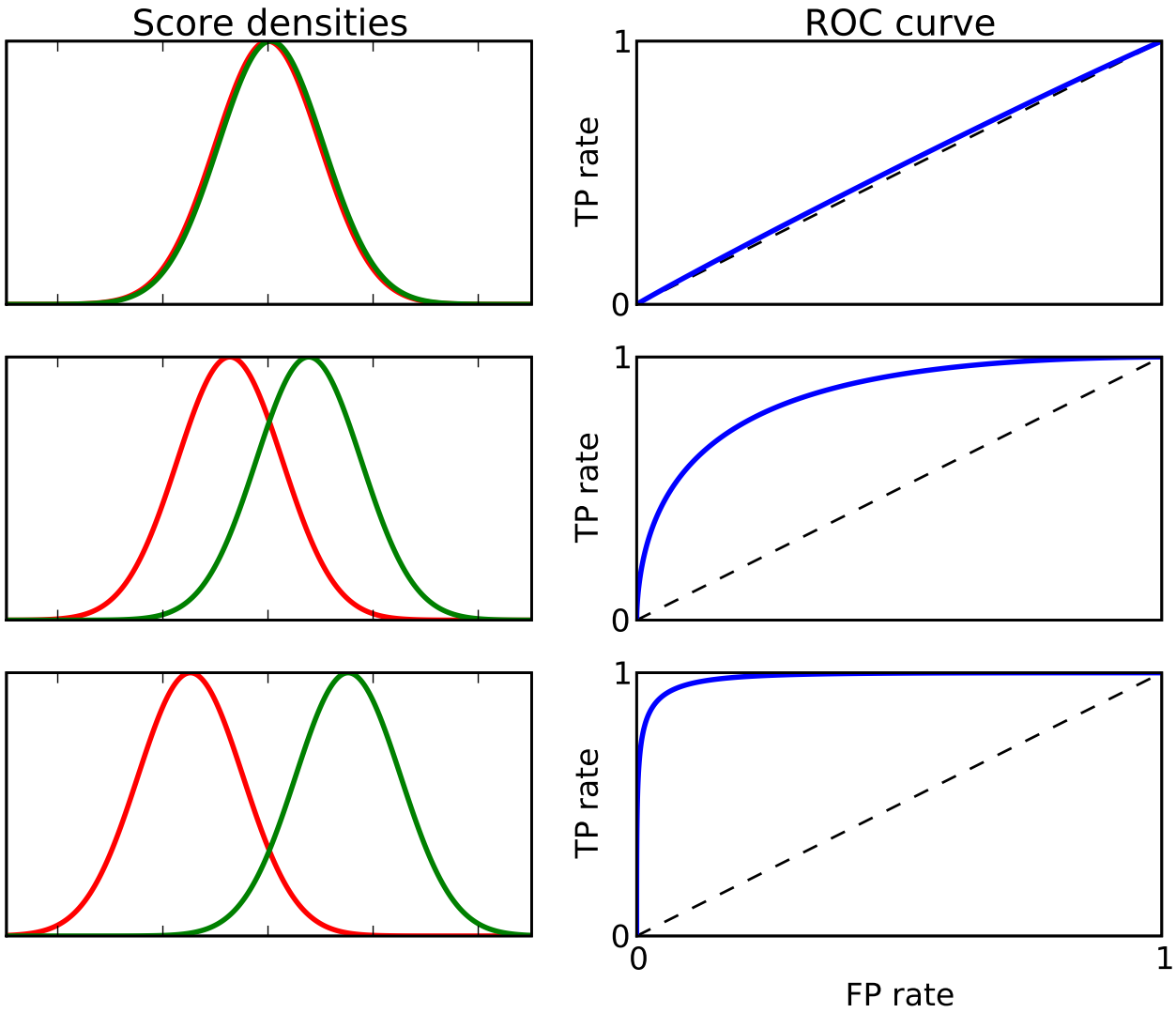
- aggregate performance over all thresholds
- arbitrary weighting of two error types (FN and FP)

$$\textbf{True positive rate (tpr)} = \frac{\#\text{True positives}}{\#\text{Positives}}$$

$$\textbf{False positive rate (fpr)} = \frac{\#\text{False positives}}{\#\text{Negatives}}$$

Illustration of ROC Curves

ROC curve: plot of true positive rate (tpr) against false positive rate (fpr) for different thresholds.



ROC Curves

ROC pros and cons

- curve is independent of prior class probabilities
- curve is independent of cost values
- fails to address the “real” issue: expected cost (measure, view, minimize, compare, etc.)

See: ICML’04 tutorial [Flach, 2004], intro paper [Fawcett, 2006]

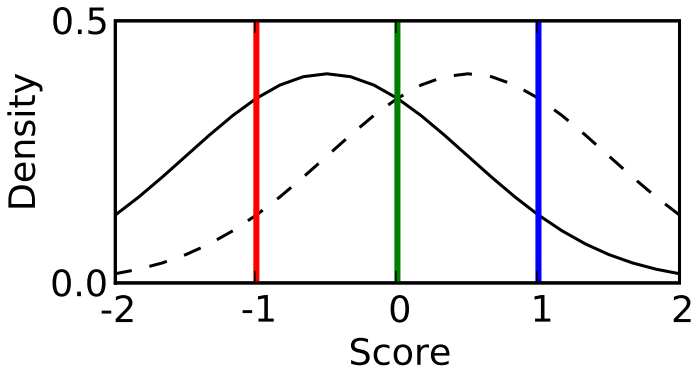
Cost Curves
[Drummond and Holte, 2000],[Drummond and Holte, 2006]

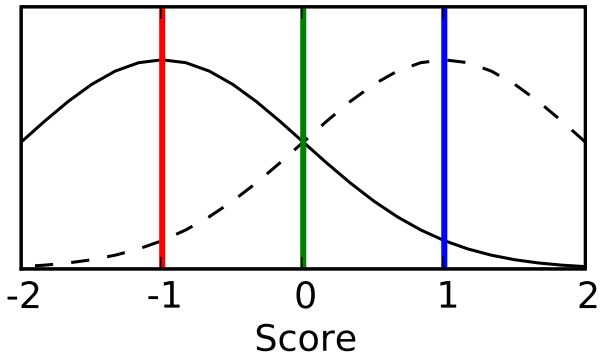
- Operating conditions:
  - misclassification costs ($c_{+|-}, c_{-|+}$)
  - prior probabilities (p_+, p_-).
- Expected cost = $p_- \cdot fpr \cdot c_{+|-} + p_+ \cdot (1 - tpr) \cdot c_{-|+}$.
- Operating point: $w = \frac{p_+ \, c_{-|+}}{p_+ \, c_{-|+} + p_- \, c_{+|-}} \in [0, 1]$
- Normalized cost: $(1 - w)fpr + w(1 - tpr)$.
- Given w, we choose the pair (fpr, tpr) from the ROC curve that minimizes the normalized cost.

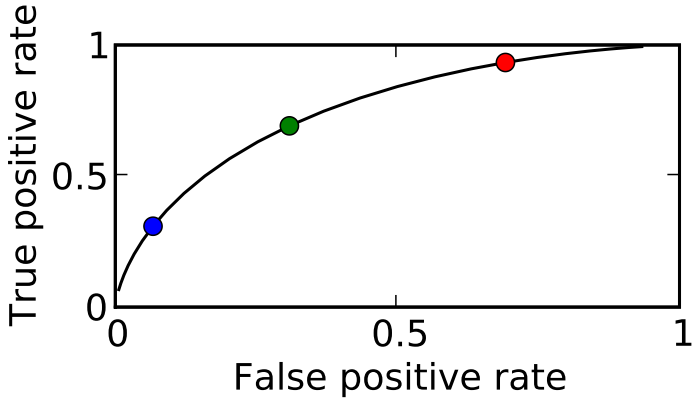
$$C(w) \quad = \quad \min_{(fpr,tpr) \in ROC} (1 - w)fpr + w(1 - tpr)$$

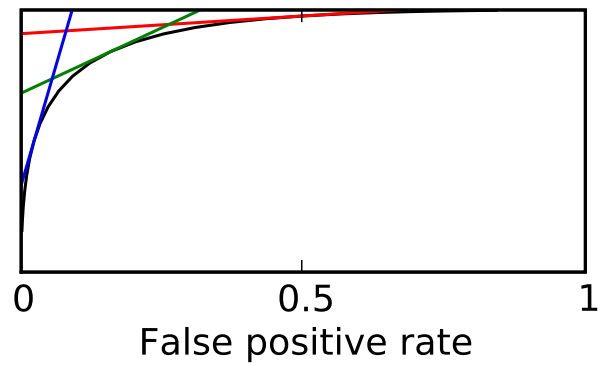
- Cost curve: plot of $C(w)$ against $w \in [0, 1]$

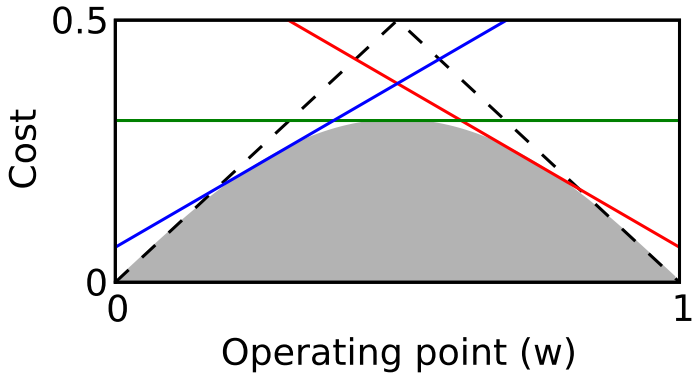
From ROC to Cost

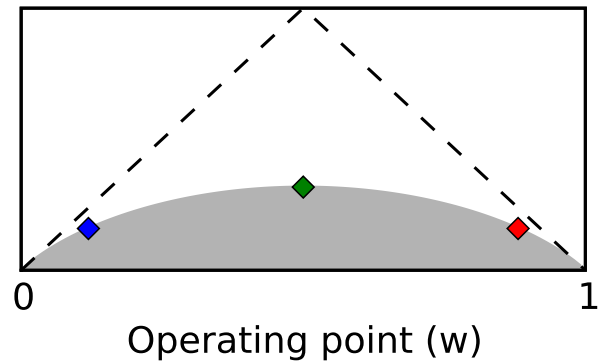






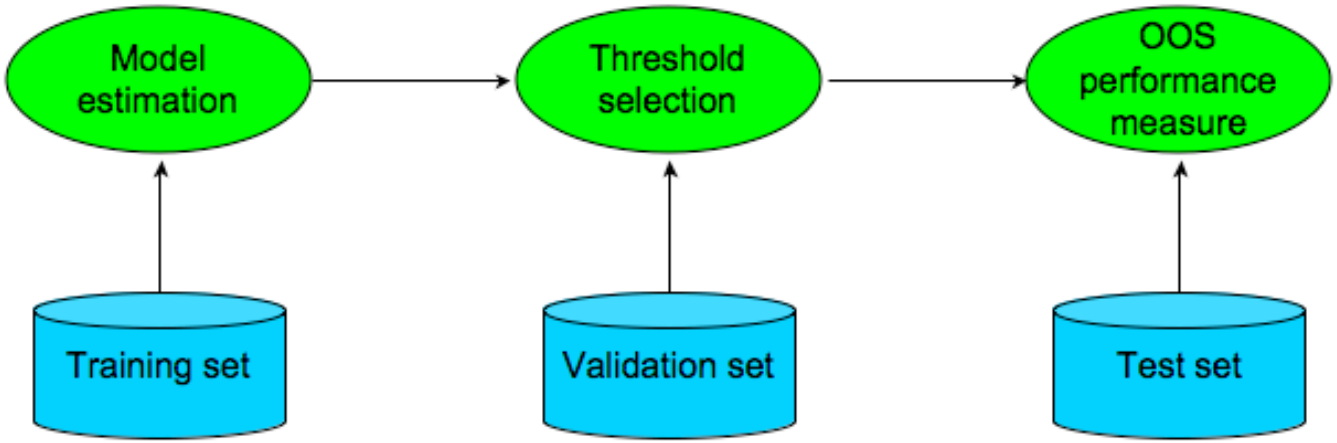






Out-of sample performance measure

- Drawing cost curve involves threshold optimization
- Must be conducted using validation set *disjoint* from test set.



- Performance distribution from single test set ?
- “Empirical” bootstrap: take samples of the test set, with replacement
- Exact bootstrap: analytic derivation for an infinite number of samples

C.I. for a single classifier's cost curve

- n: test set size
- $n^+(n^-)$: # positive (negative) instances in test set
- With prior probabilities p_+, p_- (and costs) fixed, n^+ and n^- are constant for all samples: *stratified sampling*.
- $n_t^+(n_t^-)$: # positive (negative) instances in test set with $s \geq t = t(w)$.
- $N_t^+(N_t^-)$: r.v. for # positive (negative) instances, in a given sample, with $s \geq t$.
- $TP_t = N_t^+/n^+$, $FP_t = N_t^-/n^-$
- $N_t^+ \sim \text{Bin}(n_t^+/n^+, n^+), N_t^- \sim \text{Bin}(n_t^-/n^-, n^-)$

$$\begin{aligned} C_t &= w(1 - TP_t^+) + (1-w)FP_t^- \\ E[C_t] &= w(1 - n_t^+/n^+) + (1-w)n_t^-/n^- \\ Var[C_t] &= w^2 n_t^+/n^+(1 - n_t^+/n^+) + (1-w)^2 n_t^-/n^-(1 - n_t^-/n^-) \end{aligned}$$

C.I. for difference between two cost curves

- scores of two classifiers are dependent
- thresholds may have different meanings $t_1 = t_1(w), t_2 = t_2(w)$.
- examples with $s_1 \geq t_1,\ s_2 \geq t_2$ have no effect on cost difference
- $n^+_{t_1}$: # positive instances with $s_1 \geq t_1,\ s_2 < t_2$
- $n^+_{t_2}, n^-_{t_1}, n^-_{t_2}$: defined similarly
- $N^+_{t_1}, N^+_{t_2}, N^-_{t_1}, N^-_{t_2}$: corresponding r.v.
- $(N^+_{t_1}, N^+_{t_2}) = \text{Mult}(p^+_{t_1}, p^+_{t_2}, n^+),\ p^+_{t_1} = n^+_{t_1}/n^+, p^+_{t_2} = n^+_{t_2}/n^+$

$$
\begin{aligned}
\Delta C_{t_1,t_2} &= C_{t_2} - C_{t_1} = w(TP^+_{t_1} - TP^+_{t_2}) + (1-w)(FP^-_{t_2} - FP^-_{t_1}) \\
E[\Delta C_{t_1,t_2}] &= w(p^+_{t_1} - p^+_{t_2}) + (1-w)(p^-_{t_2} - p^-_{t_1}) \\
Var[\Delta C_{t_1,t_2}] &= w^2[p^+_{t_1} + p^+_{t_2} - (p^+_{t_1} - p^+_{t_2})^2]/n^+ \\
&\quad (1-w)^2[p^-_{t_1} + p^-_{t_2} - (p^-_{t_1} - p^-_{t_2})^2]/n^-
\end{aligned}
$$

Stratified vs “Full” sampling

- Stratified sampling:
  - draw samples independently from two classes
  - cost distribution, given fixed operating point
- Full sampling:
  - draw samples from whole test set
  - cost distribution, given fixed costs but binomial distribution of class proportions
- Full sampling has larger variance

C.I. for a single classifier's cost curve (full sampling)

- $N^+(N^-)$: # positive (negative) instances in test set, now r.v.
- $c_{max} = \max(c_{+|-}, c_{-|+})$

$$
\begin{aligned}
C_t &= \frac{N^+ c_{-|+}(1 - TP_t^+) + N^- c_{+|-} FP_t^-}{n \cdot c_{max}} \\
E[C_t] &= E_{N^+}\{E[C_t|N^+]\} \\
&= \frac{c_{-/+}(n^+ - n_t^+) + c_{+/-} \cdot n_t^-}{n \cdot c_{\max}} \\
V[C_t] &= V_{N^+}\{E[C_t|N^+]\} + E_{N^+}\{V[C_t|N^+]\} \\
&= \frac{c_{-/+}^2 \alpha_t^+ + c_{+/-}^2 \alpha_t^- + \delta_t^2}{(n \cdot c_{\max})^2}
\end{aligned}
$$

$$
\alpha_t^+ = n_t^+ - \frac{(n_t^+)^2}{n^+},\ \alpha_t^- = n_t^- - \frac{(n_t^-)^2}{n^-},
$$

$$
\delta_t^2 = \left(c_{-/+}\frac{n^+ - n_t^+}{n^+} - c_{+/-}\frac{n_t^-}{n^-}\right)^2 \frac{n^+ \cdot n^-}{n}
$$

C.I. for difference between two cost curves (full sampling)

$$
\begin{aligned}
\Delta C_{t_1,t_2} &= \frac{c_{-/+}(N^+_{t_1} - N^+_{t_2}) + c_{+/-}(N^-_{t_2} - N^-_{t_1})}{n \cdot c_{\max}} \\
E[\Delta C_{t_1,t_2}] &= E_{N^+}\{E[\Delta C_{t_1,t_2}|N^+]\} \\
&= \frac{c_{-/+}(n^+_{t_1} - n^+_{t_2}) + c_{+/-} \cdot (n^-_{t_2} - n^-_{t_1})}{n \cdot c_{\max}} \\
V[\Delta C_{t_1,t_2}] &= V_{N^+}\{E[\Delta C_{t_1,t_2}|N^+]\} \\
&\quad + E_{N^+}\{V[\Delta C_{t_1,t_2}|N^+]\} \\
&= \frac{c^2_{-/+}\alpha^+_{t_1,t_2} + c^2_{+/-}\alpha^-_{t_1,t_2} + \delta^2_{t_1,t_2}}{(n \cdot c_{\max})^2}
\end{aligned}
$$

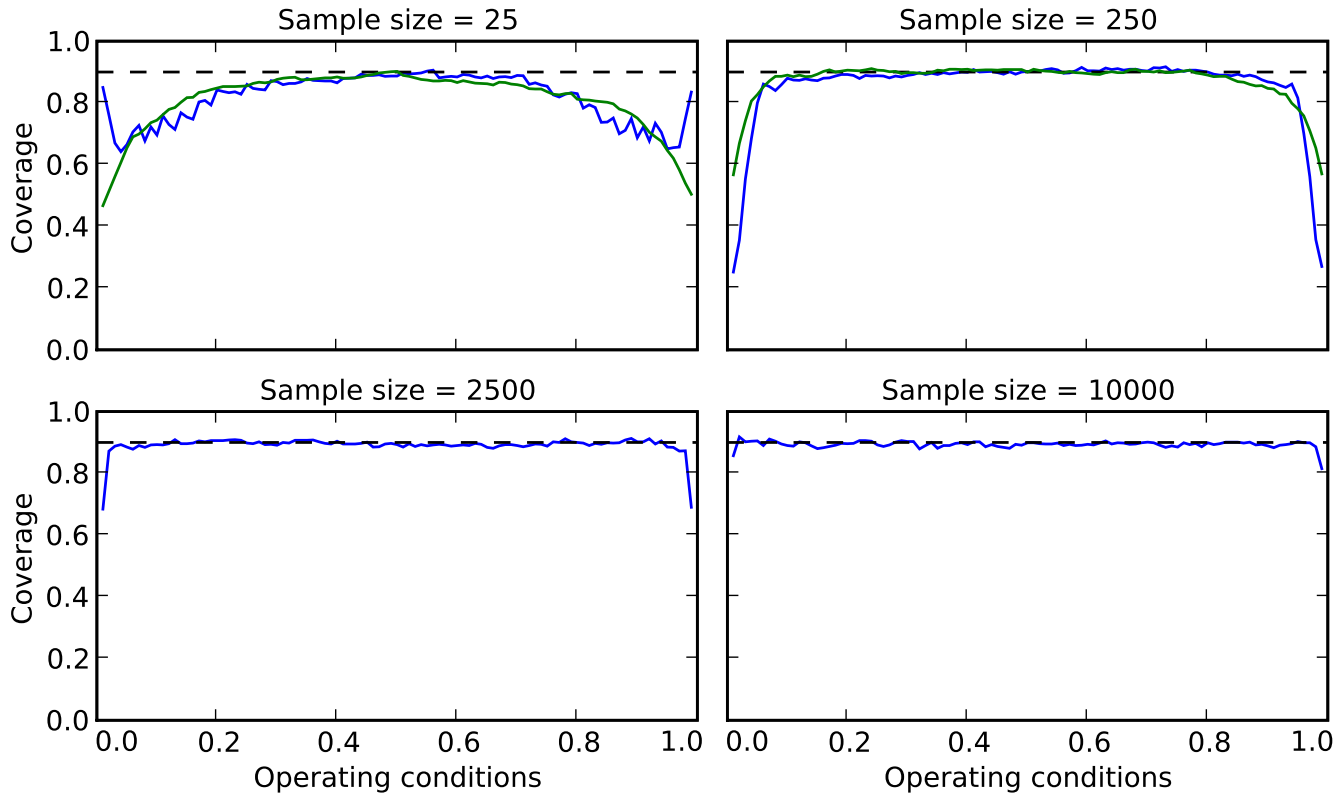
$$
\alpha^+_{t_1,t_2} = n^+_{t_1} + n^+_{t_2} - \frac{(n^+_{t_1} - n^+_{t_2})^2}{n^+}, \alpha^-_{t_1,t_2} = n^-_{t_1} + n^-_{t_2} - \frac{(n^-_{t_1} - n^-_{t_2})^2}{n^-},
$$

$$
\delta^2_{t_1,t_2} = \left(c_{-/+}\frac{n^+_{t_1} - n^+_{t_2}}{n^+} - c_{+/-}\frac{n^-_{t_2} - n^-_{t_1}}{n^-} \right)^2 \frac{n^+ \cdot n^-}{n}
$$

Simulations (one curve)

- Scores of positive instances $\sim N(\mu = 3, \sigma = 3)$
- Scores of negative instances $\sim N(\mu = -3, \sigma = 3)$
- Thresholds set to cost minimizing according to distribution
- Samples of 25, 250, 2500 and 10000 drawn to compute p.e.b.c.i.
- 1000 simulations
- Coverage = proportion of simul. with true curve included in C.I.
- $\alpha = 10\%$, i.e. 90% C.I.
- $w \in \{0.01, 0.02, \ldots, 0.99\}$

Simulations (one curve - stratified sampling)

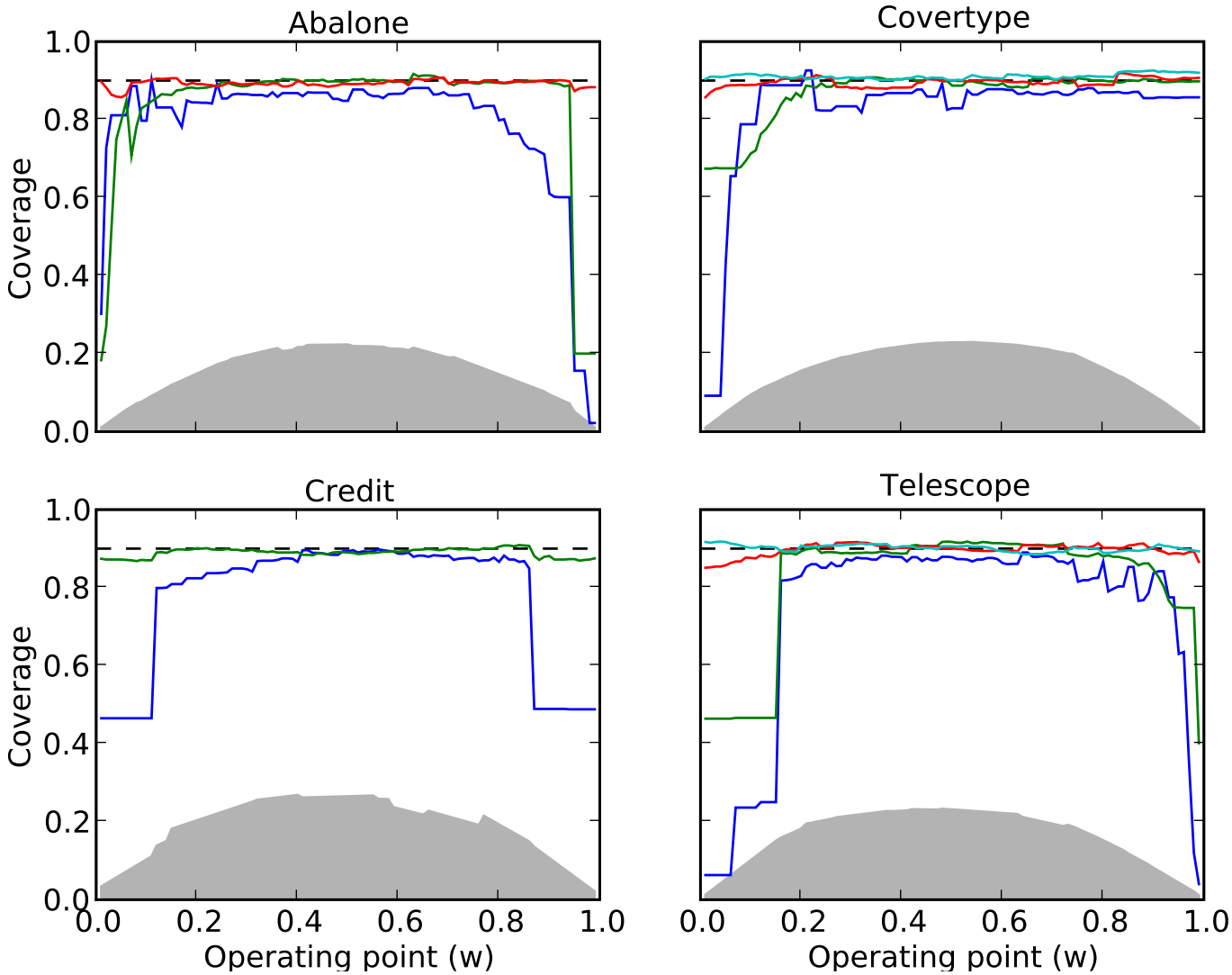


UCI experiments (one curve)

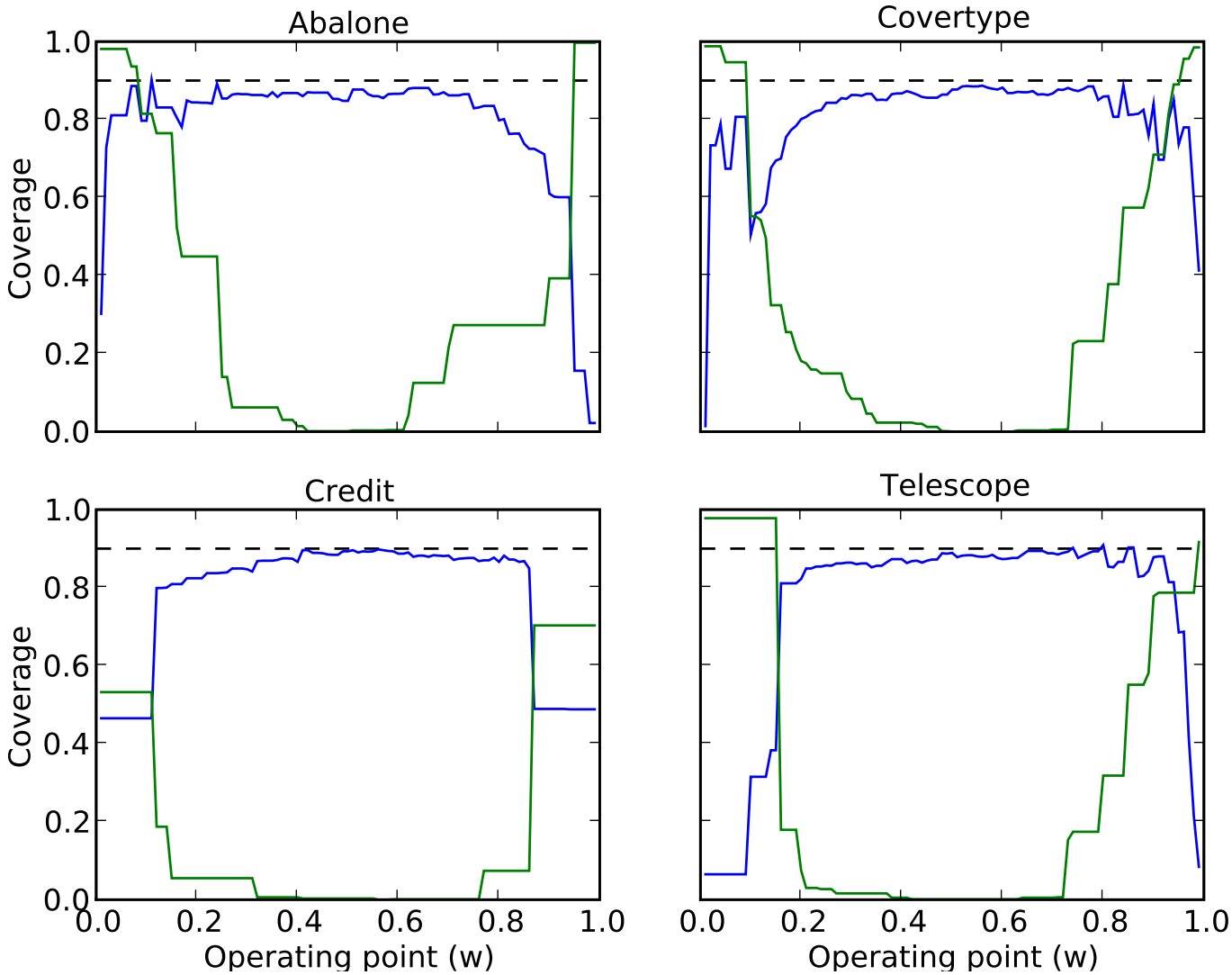
Dataset	Train	Valid	Test	(perc. pos.)
Abalone	1000	1000	3177	(50%)
Covertype	5000	5000	485141	(57%)
Credit (german)	500	250	250	(69%)
Telescope (magic)	1000	1000	17020	(65%)

- Logistic regression models
- Entire test set used to compute "true" cost curve
- Samples of 25, 250, 2500 and 10000 drawn to compute p.e.b.c.i.
- 1000 simulations
- Coverage = proportion of simul. with true curve included in C.I.
- $\alpha = 10\%$, i.e. 90% C.I.
- $w \in \{0.01, 0.02, \ldots, 0.99\}$

UCI experiments (one curve - stratified sampling)



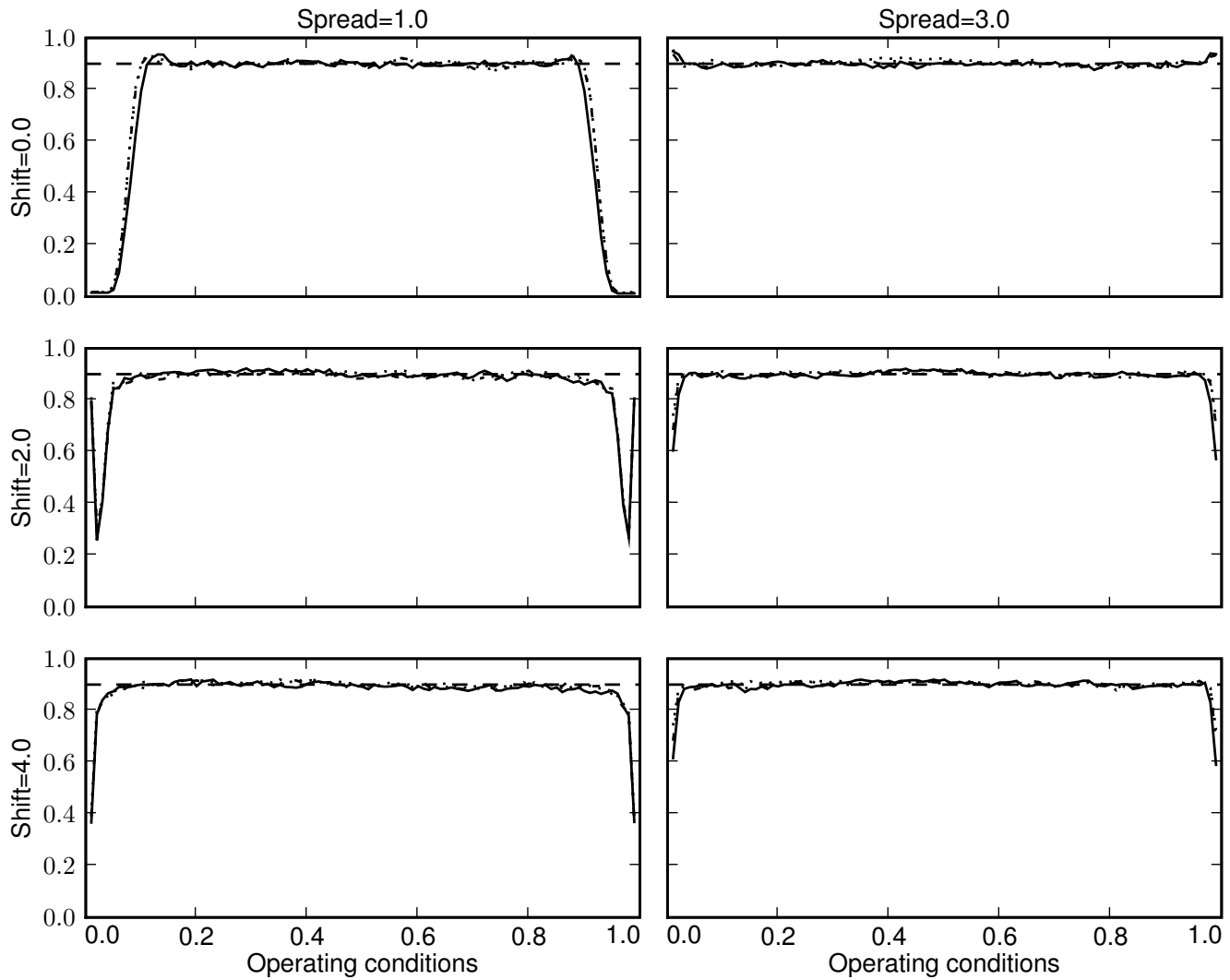
UCI experiments (one curve - stratified sampling)



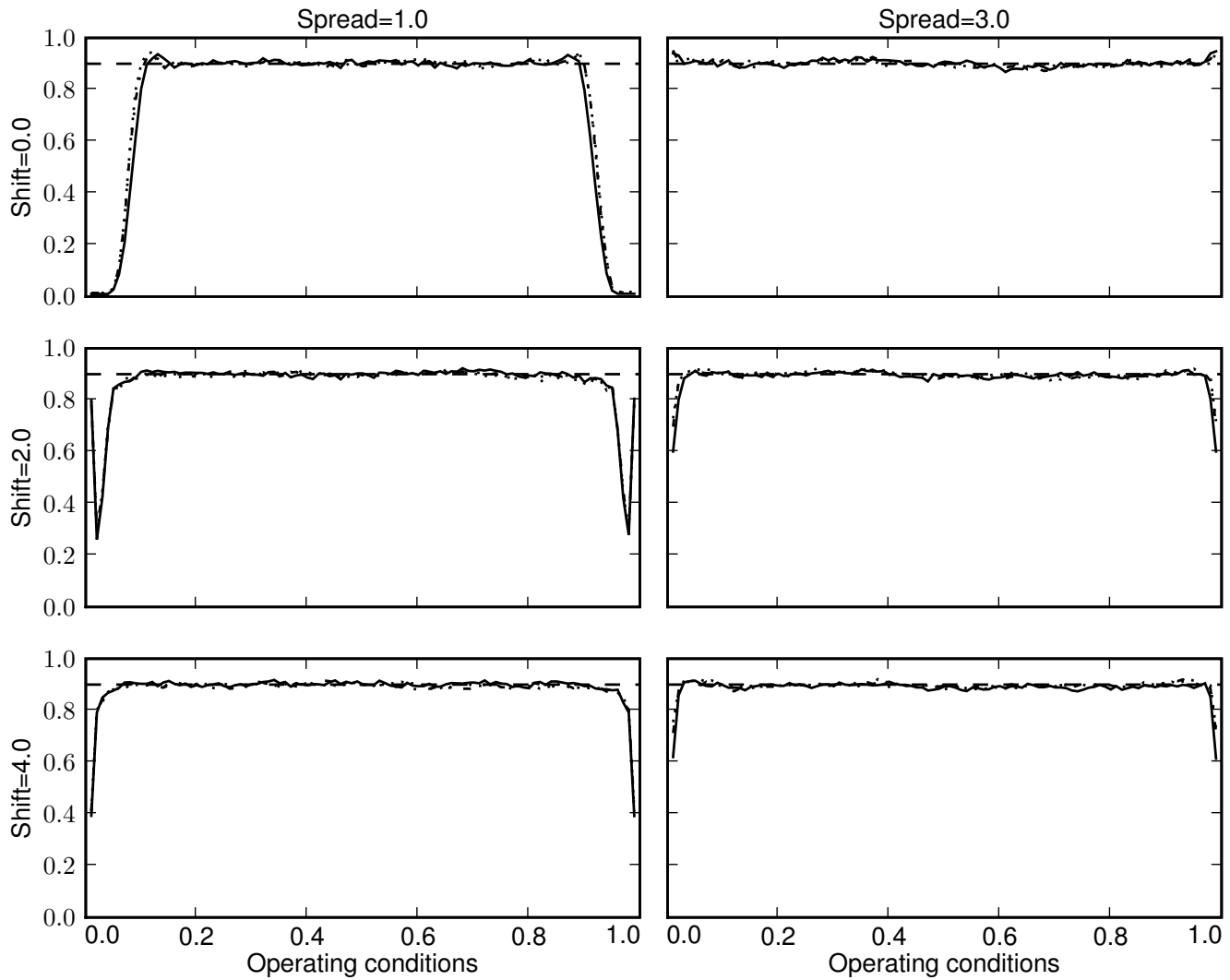
Simulations (two curves)

- Scores of positive instances, 1st model: $\sim N(\mu = \theta, \sigma = 3)$
- Scores of positive instances, 2nd model: $\sim N(\mu = \theta + \delta, \sigma = 3)$
- Scores of negative instances $\sim N(\mu = -\theta, \sigma = 3)$
- Spread: $\theta = 1.0, 3.0$
- Shift: $\delta = 0.0, 2.0, 4.0$
- Score correlation $\rho = 0.3, 0.6, 0.9$
- Thresholds set to cost minimizing according to distribution
- Sample size: 10000
- 1000 simulations
- $\alpha = 10\%$, i.e. 90% C.I.
- $w \in \{0.01, 0.02, \ldots, 0.99\}$

Simulations (two curves - stratified sampling)



Simulations (two curves - full sampling)



Discussion

- Cost curves are an excellent visualization tool of the “true” target: expected cost
- Provided means to compute confidence intervals of cost curves for
  - Stratified or full sampling
  - One or two curves
  - Fast: $O(n \log n)$ (once sorted, everything is linear)
- Empirical method, can not extrapolate.
- Solutions against breaks: kernels, tail distribution estimation

References


Drummond, C. and Holte, R. (2000).
Explicitly representing expected cost: an alternative to ROC representation.
In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 198–207. ACM.


Drummond, C. and Holte, R. (2006).
Cost curves: an improved method for visualizing classifier performance.
Machine Learning, 65(1):95–130.


Fawcett, T. (2006).
An introduction to ROC analysis.
Pattern Recognition Letters, 27(8):861–874.


Flach, P. (2004).
The many faces of ROC analysis in machine learning.