Evaluating predictive uncertainty

Balaji Lakshminarayanan

Based on [QCRS+06]

Motivation

- Predictive uncertainty is essential in decision making
  - Probability of cancer is 99%, Stock will increase by $10 \pm 1\%$
  - Active learning: select next training example, experimental design

Motivation

- Predictive uncertainty is essential in decision making
  - Probability of cancer is 99%, Stock will increase by $10 \pm 1\%$
  - Active learning: select next training example, experimental design
- Loss function may be unknown
  - Provide predictive uncertainty over quantity of interest

Motivation

- Predictive uncertainty is essential in decision making
  - Probability of cancer is 99%, Stock will increase by $10 \pm 1\%$
  - Active learning: select next training example, experimental design
- Loss function may be unknown
  - Provide predictive uncertainty over quantity of interest
- Approaches: Bayesian model averaging, bagging, other hacks or principles :)

Motivation

- Predictive uncertainty is essential in decision making
  - Probability of cancer is 99%, Stock will increase by $10 \pm 1\%$
  - Active learning: select next training example, experimental design
- Loss function may be unknown
  - Provide predictive uncertainty over quantity of interest
- Approaches: Bayesian model averaging, bagging, other hacks or principles :)
- How do we evaluate predictive uncertainty?
  - How do Bayesian methods fare on the empirical battleground?

Motivation (contd.)

"It is clear that the merits of Bayesian and competing approaches will not be settled by philosophical disputation, but only by demonstrations of effectiveness in practical contexts."

Motivation (contd.)

"It is clear that the merits of Bayesian and competing approaches will not be settled by philosophical disputation, but only by demonstrations of effectiveness in practical contexts."

- Radford Neal (PhD thesis)

Probabilistic predictions

- Binary Classification: $p(y_* = 1 | x_*)$
- Regression:
  - Unimodal: Gaussian with mean m_* and variance v_*
  - Multimodal: N quantiles $[q_{\alpha_1}, \ldots, q_{\alpha_N}]$ such that $p(y_* < q_{\alpha_j} | x_*) = \alpha_j$ where $0 < \alpha_j < 1$.

Multimodal posterior for regression

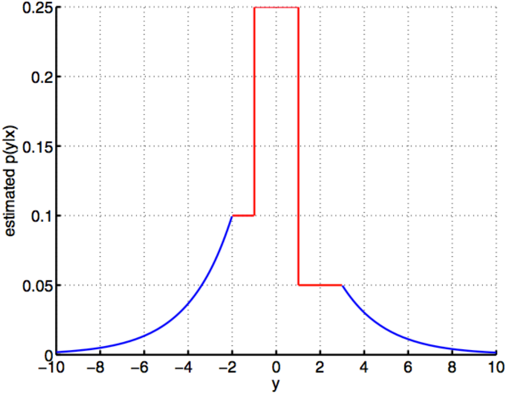


Fig. 4. Specifying the predictive density with quantiles. Example where the quantiles $q_{0.2} = -2$, $q_{0.3} = -1$, $q_{0.8} = 1$ and $q_{0.9} = 3$ are specified. The exponential tails guarantee that distribution integrates to 1.

Loss functions for classification

- Average classification error (threshold=0.5)
- Negative log probability (NLP) loss

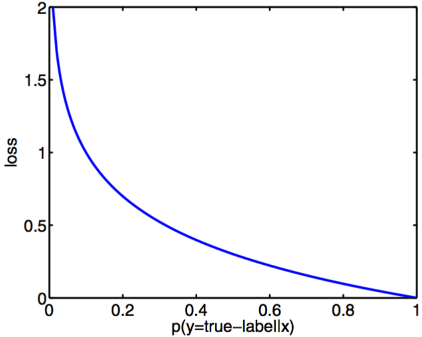


Fig. 5. NLP loss when predicting the class of a single test point that actually belongs to class "+1". Observe how the loss goes to infinity as the model becomes increasingly certain that the point belongs to the wrong class.

- LIFT loss, calibration curve, Brier score, ...

Loss functions for regression

- normalized Mean squared error (nMSE)
- Negative log probability density (NLPD)

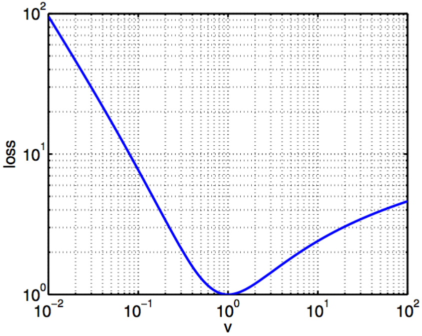


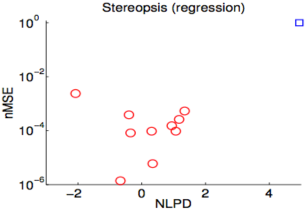
Fig. 7. NLPD loss (up to a constant) incurred when predicting at a single point with a Gaussian predictive distribution. In the figure we have fixed $\|t_i - m_i\|^2 = 1$ and show how the loss evolves as we vary the predictive variance v_i. The optimal value of the predictive variance is equal to the actual squared error given the predictive mean.

Discussion about losses

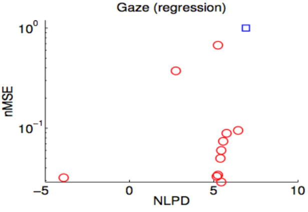
- Log loss for classification: Infinite penalty too strong? Strongly discourages overconfident wrong predictions
- NLPD can be "gamed" when same value is repeated multiple times (eg. data is ordinal rather than real-valued)
- Other metrics: Mutual information, AUC
  - Aggregate vs point wise metrics?
  - Account only for relative degrees of belief, sometimes we care about absolute values and not just the ordering

Results

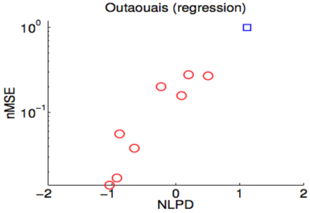
Regression



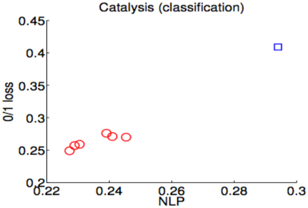
(a)

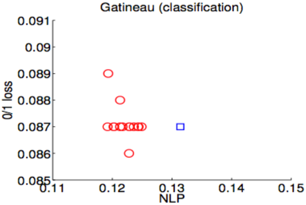


(b)

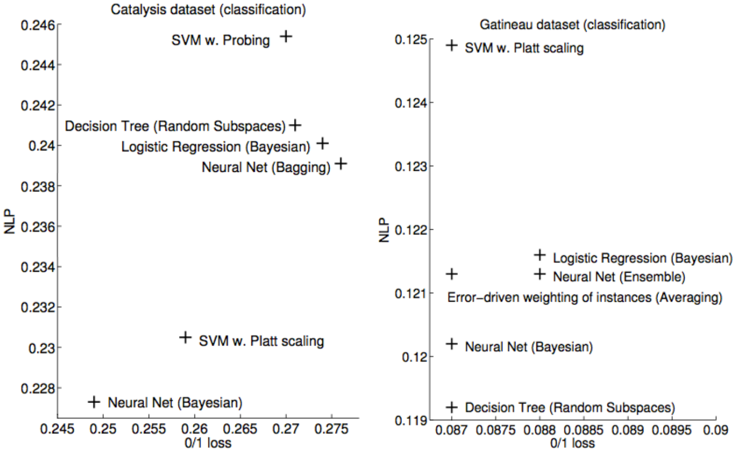


Classification





Classification



Classification: Catalysis

Catalysis (Classification)

Method	NLP	01L	Author
Bayesian NN	0.2273	0.249	**Neal, R**
< Bayesian NN	0.2289	0.257	**Neal, R**
SVM + Platt	0.2305	0.259	**Chapelle, O**
> Bagged R-MLP	0.2391	0.276	**Cawley, G**
> Bayesian Logistic Regression	0.2401	0.274	**Neal, R**
Feat Sel + Rnd Subsp + Dec Trees	0.2410	0.271	**Chawla, N**
Probing SVM	0.2454	0.270	Zadrozny, B & Langford, J
baseline: class frequencies	**0.2940**	**0.409**	

(NLP: average negative log probability, 01L: average zero-one loss)

Classification: Gatineau

Gatineau (Classification)

Method	NLP	01L	Author
Feat Sel + Rnd subsp + Dec Trees	0.1192	0.087	**Chawla, N**
Feat Sel + Bagging + Dec Trees	0.1193	0.089	**Chawla, N**
Bayesian NN	0.1202	0.087	**Neal, R**
< Bayesian NN	0.1203	0.087	**Neal, R**
Simple ANN Ensemble	0.1213	0.088	Ohlsson, M
EDWIN	0.1213	0.087	Eisele, A
> Bayesian Logistic Regression	0.1216	0.088	**Neal, R**
> ANN with L1 penalty	0.1217	0.087	Delalleau, O
> CCR-MLP	0.1228	0.086	**Cawley, G**
Rnd Subsp + Dec Trees	0.1228	0.087	**Chawla, N**
Bagging + Dec Trees	0.1229	0.087	**Chawla, N**
> R-MLP	0.1236	0.087	**Cawley, G**
Probing J48	0.1243	0.087	Zadrozny, B & Langford, J
> Bagged R-MLP (small)	0.1244	0.087	**Cawley, G**
SVM + Platt	0.1249	0.087	**Chapelle, O**
baseline: class frequencies	**0.1314**	**0.087**	

(NLP: average negative log probability, 01L: average zero-one loss)

Regression: Stereopsis

Stereopsis (Regression)

Method	NLPD	nMSE	Author
Mixture of Bayesian Neural Nets	-2.077	2.38e-3	**Snelson & Murray**
Compet Assoc Nets + Cross Val	-0.669	1.39e-6	**Kurogi, S et al**
> Mixt of LOOHKRR Machines	-0.402	3.86e-4	**Cawley, G**
> Gaussian Process Regression	-0.351	8.25e-5	**Chapelle, O**
> Inflated Var MLP Committee	0.309	9.59e-5	**Cawley, G**
KRR + Regression on the variance	0.342	9.60e-5	**Chapelle, O**
< Hybrid: Neural Net	0.940	1.52e-4	Lewandowski, A
Mixture Density Network Ensemble	1.171	2.62e-4	Carney, M
baseline: empirical Gaussian	**4.94**	**1.002**	
Modelling the experimental setting	209.4	2.49e-4	**Kohonen & Suomela**

(NPLD: negative log predictive density, nMSE: normalized mean squared error)

Regression: Gaze

Gaze (Regression)

Method	NLPD	nMSE	Author
Compet Assoc Nets + Cross Val	-3.907	0.032	**Kurogi, S et al**
LLR Regr + Resid Regr + Int Spikes	2.750	0.374	**Kohonen & Suomela**
> LOOHKRR	5.180	0.033	**Cawley, G**
> Heteroscedastic MLP Committee	5.248	0.034	**Cawley, G**
Gaussian Process regression	5.250	0.675	Csató, L
KRR + Regression on the variance	5.395	0.050	**Chapelle, O**
< Neural Net	5.444	0.029	Lewandowski, A
Rand Forest with OB enhancement	5.445	0.060	Van Matre, B
NeuralBAG and EANN	5.558	0.074	Carney, M
Mixture Density Network Ensemble	5.761	0.089	Carney, M
baseline: empirical Gaussian	**6.91**	**1.002**	

Regression: Outaousis

Outaouais (Regression)

Method	NLPD	nMSE	Author
> Sparse GP method	-1.037	0.014	Keerthi & Chu
> Gaussian Process regression	-0.921	0.017	Chu, Wei
Classification + Nearest Neighbour	-0.880	0.056	**Kohonen, J**
Compet Assoc Nets + Cross Val	-0.648	0.038	**Kurogi S et al**
> Small Heteroscedastic MLP	-0.230	0.201	**Cawley, G**
Gaussian Process regression	0.090	0.158	Csató, L
Mixture Density Network Ensemble	0.199	0.278	Carney, M
NeuralBAG and EANN	0.505	0.270	Carney, M
baseline: empirical Gaussian	**1.115**	**1.000**	

Summary

- Defining good losses for probabilistic predictions is hard
  - How to encourage "honest" (loss-indepenent) predictive distributions?
  - Apply several losses that have contradictory properties
- Datasets and losses should not be chosen separately, since some losses are inappropriate for evaluating performance on certain problems.
  - log loss for regression is not appropriate when the same target occurs more than once
- Bayesian methods aren't the only competitive methods; non-Bayesian approaches, (like regression on the variance [CTC06]), did also perform very well.

Gavin C Cawley, Nicola LC Talbot, and Olivier Chapelle.
Estimating predictive variances with kernel ridge regression.
In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 56–77. Springer, 2006.

Joaquin Quinonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf.
Evaluating predictive uncertainty challenge.
In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 1–27. Springer, 2006.