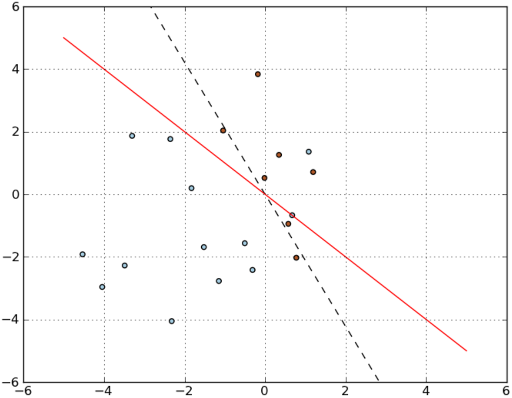
Learning with Marginalized Corrupted Features

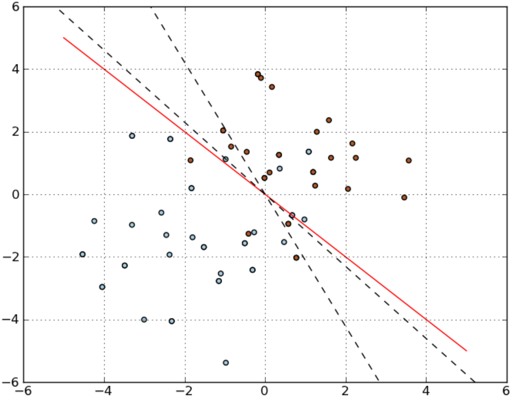
L. van der Maaten, M. Chen, S. Tyree, K. Weinberger

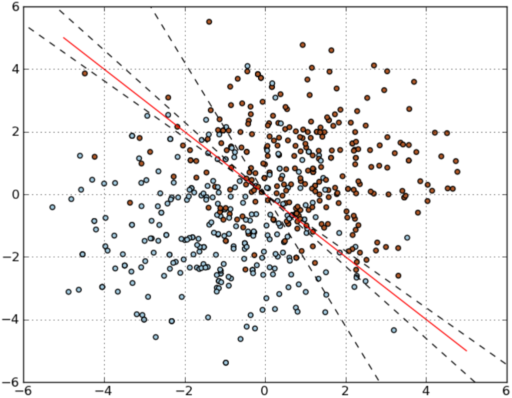
ICML 2013

Jan Gasthaus

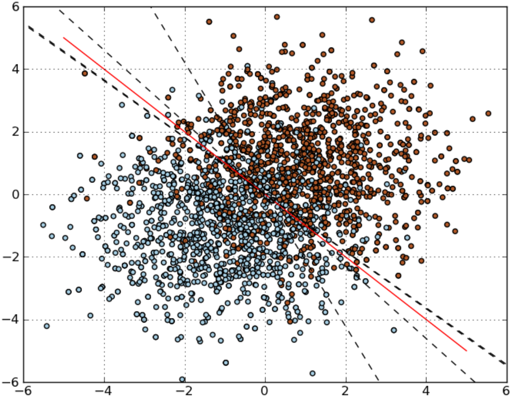
Tea talk
April 11, 2013

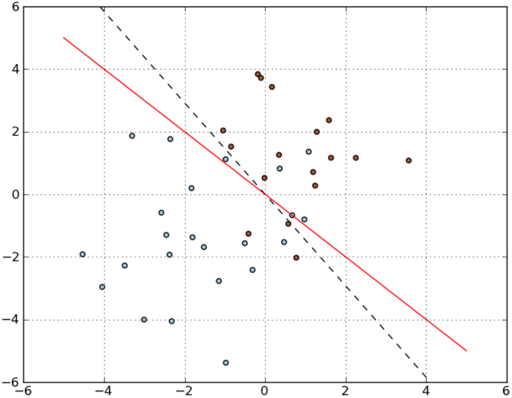

6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

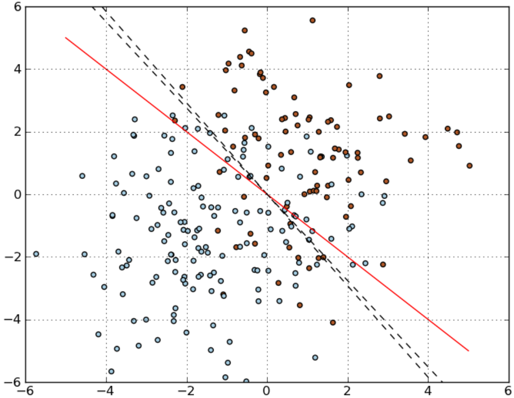

6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

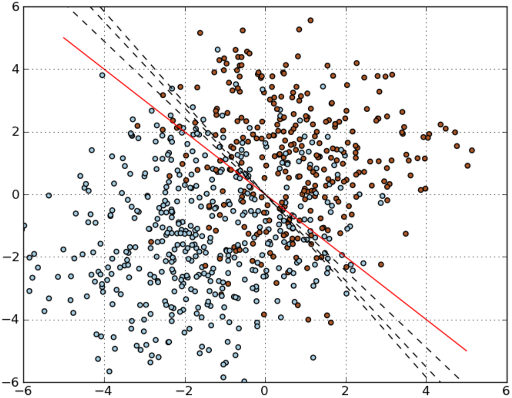


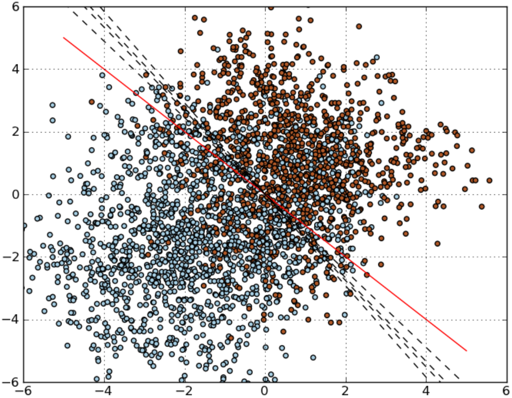
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

Data Augmentation

Secret 4: lots of jittering, mirroring, and color perturbation of the original images generated on the fly to increase the size of the training set

Yann LeCun on Google+ about Alex Krizhevsky's ImageNet results

Main Idea

- Old idea: create artificial additional training data by corrupting it with “noise”

Main Idea

- Old idea: create artificial additional training data by corrupting it with “noise”
- One easy way to incorporate domain knowledge (e.g. possible transformations)

Main Idea

- Old idea: create artificial additional training data by corrupting it with “noise”
- One easy way to incorporate domain knowledge (e.g. possible transformations)
- But: additional training data $\Longrightarrow$ additional computation
- **Idea:** Corrupt with known ExpFam noise and integrate it out

Explicit vs. Implicit Corruption

- Explicit corruption: Take training set $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ and corrupt it M times

$$\mathcal{L}(\tilde{D}, \Theta) = \sum_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} L(\tilde{\mathbf{x}}_{nm}, y_n, \Theta)$$

with $\mathbf{x}_{nm} \sim p(\tilde{\mathbf{x}}_{nm} | \mathbf{x}_n)$.

Explicit vs. Implicit Corruption

- Implicit corruption: Minimize the expected value of the loss under $p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)$:

$$\mathcal{L}(D, \Theta) = \sum_{n=1}^{N} \mathbb{E}\left[L(\tilde{\mathbf{x}}_n, y_n, \Theta)\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)}$$

i.e. replace the empirical average with the expectation.

Wait a second . . .

- This is *so* obvious that it must have been done before . . .

Wait a second . . .

- This is *so* obvious that it must have been done before . . .
  - *Vicinal Risk Minimization*, Chapelle, Weston, Bottou, & Vapnik, NIPS 2000

$$R_{vic}(f) = \int \ell(f(\mathbf{x}), y)\, dP_{est}(\mathbf{x}, y) = \frac{1}{n}\sum_{i=1}^{n} \int \ell(f(\mathbf{x}), y_i) dP_{\mathbf{x}_i}(\mathbf{x})$$

Wait a second . . .

- This is *so* obvious that it must have been done before . . .
  - *Vicinal Risk Minimization*, Chapelle, Weston, Bottou, & Vapnik, NIPS 2000

$$R_{vic}(f) = \int \ell(f(\mathbf{x}), y)\, dP_{est}(\mathbf{x}, y) = \frac{1}{n}\sum_{i=1}^{n} \int \ell(f(\mathbf{x}), y_i) dP_{\mathbf{x}_i}(\mathbf{x})$$

- Explicitly only consider the case of Gaussian noise distributions

Quadratic Loss

Quadratic loss. Assuming[2] a label variable $y \in \{-1, +1\}$, the expected value of the quadratic loss under corrupting distribution $p(\tilde{\mathbf{x}}|\mathbf{x})$ is given by:

$$\mathcal{L}(\mathcal{D}; \mathbf{w}) = \sum_{n=1}^{N} \mathbb{E}\left[\left(\mathbf{w}^{\mathrm{T}}\tilde{\mathbf{x}}_n - y_n\right)^2\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)}$$

$$= \mathbf{w}^{\mathrm{T}}\left(\sum_{n=1}^{N} \mathbb{E}[\tilde{\mathbf{x}}_n]\mathbb{E}[\tilde{\mathbf{x}}_n]^{\mathrm{T}} + V[\tilde{\mathbf{x}}_n]\right)\mathbf{w}$$

$$- 2\left(\sum_{n=1}^{N} y_n \mathbb{E}[\tilde{\mathbf{x}}_n]\right)^{\mathrm{T}} \mathbf{w} + N, \quad (3)$$

$$\mathbf{w}^* = \left(\sum_{n=1}^{N} \mathbb{E}[\tilde{\mathbf{x}}_n]\mathbb{E}[\tilde{\mathbf{x}}_n]^{\mathrm{T}} + V[\tilde{\mathbf{x}}_n]\right)^{-1}\left(\sum_{n=1}^{N} y_n \mathbb{E}[\tilde{\mathbf{x}}_n]\right)$$

Quadratic Loss

Distribution	PDF	$\mathbb{E}[\tilde{x}_{nd}]_{p(\tilde{x}_{nd}\|x_{nd})}$	$V[\tilde{x}_{nd}]_{p(\tilde{x}_{nd}\|x_{nd})}$
Blankout noise	$p(\tilde{x}_{nd} = 0) = q_d$ $p(\tilde{x}_{nd} = \frac{1}{1-q_d} x_{nd}) = 1 - q_d$	x_{nd}	$\frac{q_d}{1-q_d} x_{nd}^2$
Gaussian noise	$p(\tilde{x}_{nd}\|x_{nd}) = \mathcal{N}(\tilde{x}_{nd}\|x_{nd}, \sigma^2)$	x_{nd}	σ^2
Laplace noise	$p(\tilde{x}_{nd}\|x_{nd}) = Lap(\tilde{x}_{nd}\|x_{nd}, \lambda)$	x_{nd}	$2\lambda^2$
Poisson noise	$p(\tilde{x}_{nd}\|x_{nd}) = Poisson(\tilde{x}_{nd}\|x_{nd})$	x_{nd}	x_{nd}

Exponential Loss

Exponential loss. The expected value of the exponential loss under corruption model $p(\tilde{\mathbf{x}}|\mathbf{x})$ is:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \mathbb{E}\left[e^{-y_n \mathbf{w}^{\mathrm{T}} \tilde{\mathbf{x}}_n}\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)}$$

$$= \sum_{n=1}^{N} \prod_{d=1}^{D} \mathbb{E}\left[e^{-y_n w_d \tilde{x}_{nd}}\right]_{p(\tilde{x}_{nd}|x_{nd})}, \qquad (4)$$

Logistic Loss

Logistic loss. In the case of the logistic loss, the solution to (2) cannot be computed in closed form. Instead, we derive an upper bound, which can be minimized as a surrogate loss:

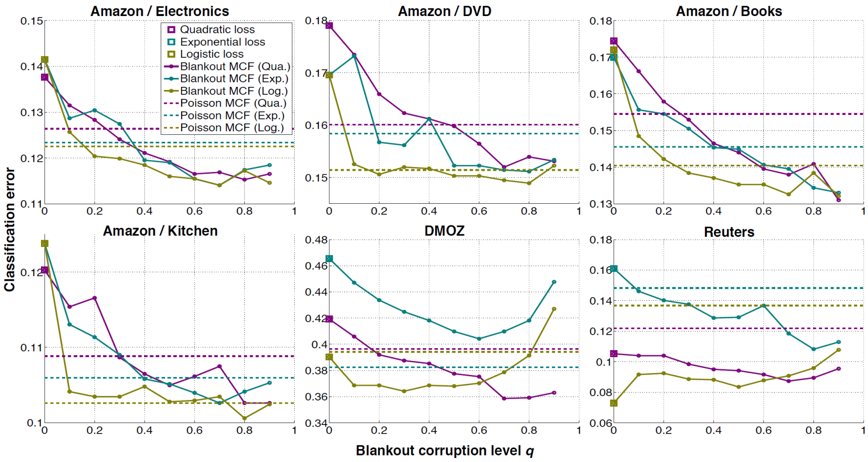
$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \mathbb{E}\left[\log\left(1+e^{-y_n \mathbf{w}^{\mathrm{T}}\tilde{\mathbf{x}}_n}\right)\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)}$$

$$\leq \sum_{n=1}^{N} \log\left(1+\prod_{d=1}^{D}\mathbb{E}\left[e^{-y_n w_d \tilde{x}_{nd}}\right]_{p(\tilde{x}_{nd}|x_{nd})}\right). \quad (5)$$

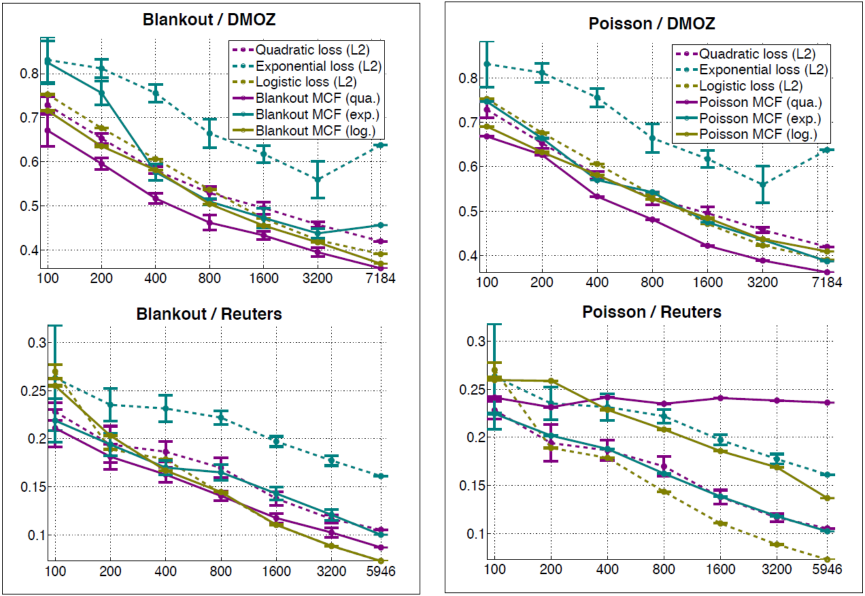
MGFs

Distribution	$\mathbb{E}[\exp(-y_n w_d \tilde{x}_{nd})]_{p(\tilde{x}_{nd}\|x_{nd})}$
Blankout noise	$q_d + (1-q_d)\exp(-y_n w_d \frac{1}{1-q_d} x_{nd})$
Gaussian noise	$\exp(-y_n w_d x_{nd} + \frac{1}{2}\sigma^2 y_n^2 w_d^2)$
Laplace noise	$(1-\lambda^2 y_n^2 w_d^2)^{-1}\exp(-y_n w_d x_{nd})$
Poisson noise	$\exp(x_{nd}(\exp(-y_n w_d) - 1))$

Results

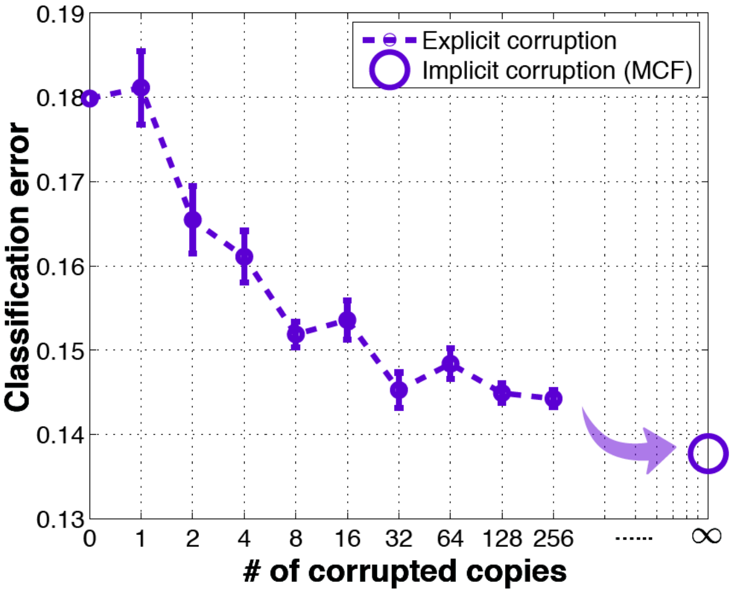

Amazon / Electronics
Amazon / DVD
Amazon / Books
Amazon / Kitchen
DMOZ
Reuters
Quadratic loss
Exponential loss
Logistic loss
Blankout MCF (Qua.)
Blankout MCF (Exp.)
Blankout MCF (Log.)
Poisson MCF (Qua.)
Poisson MCF (Exp.)
Poisson MCF (Log.)
Classification error
Blankout corruption level q

Results



Blankout / DMOZ
Quadratic loss (L2)
Exponential loss (L2)
Logistic loss (L2)
Blankout MCF (qua.)
Blankout MCF (exp.)
Blankout MCF (log.)
0.4
0.5
0.6
0.7
0.8
100
200
400
800
1600
3200
7184
Poisson / DMOZ
Quadratic loss (L2)
Exponential loss (L2)
Logistic loss (L2)
Poisson MCF (qua.)
Poisson MCF (exp.)
Poisson MCF (log.)
0.4
0.5
0.6
0.7
0.8
100
200
400
800
1600
3200
7184
Blankout / Reuters
0.1
0.15
0.2
0.25
0.3
100
200
400
800
1600
3200
5946
Poisson / Reuters
0.1
0.15
0.2
0.25
0.3
100
200
400
800
1600
3200
5946

Results



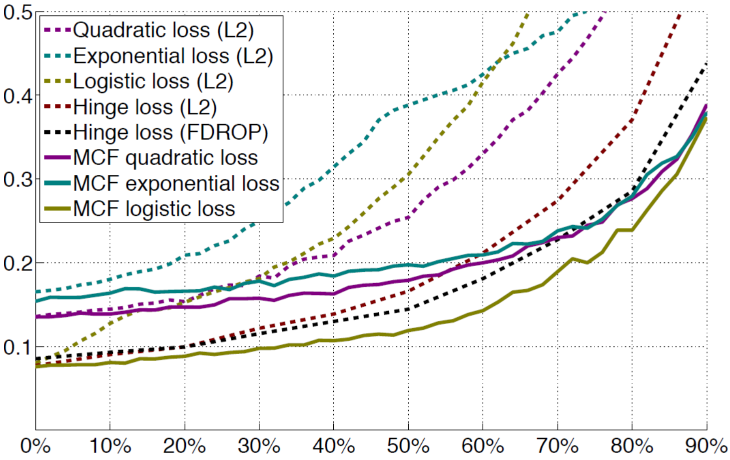
Explicit corruption
Implicit corruption (MCF)
Classification error
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0
1
2
4
8
16
32
64
128
256
......
∞
# of corrupted copies

Results

	Quadr.	Expon.	Logist.
No MCF	32.6%	39.7%	32.5%
Poisson MCF	29.1%	39.5%	30.0%
Blankout MCF	32.3%	37.9%	29.4%

Table 3. Classification errors obtained on the CIFAR-10 data set with MCF classifiers trained on simple spatial-pyramid bag-of-visual-words features.

Results


0.5
0.4
0.3
0.2
0.1
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
Quadratic loss (L2)
Exponential loss (L2)
Logistic loss (L2)
Hinge loss (L2)
Hinge loss (FDROP)
MCF quadratic loss
MCF exponential loss
MCF logistic loss