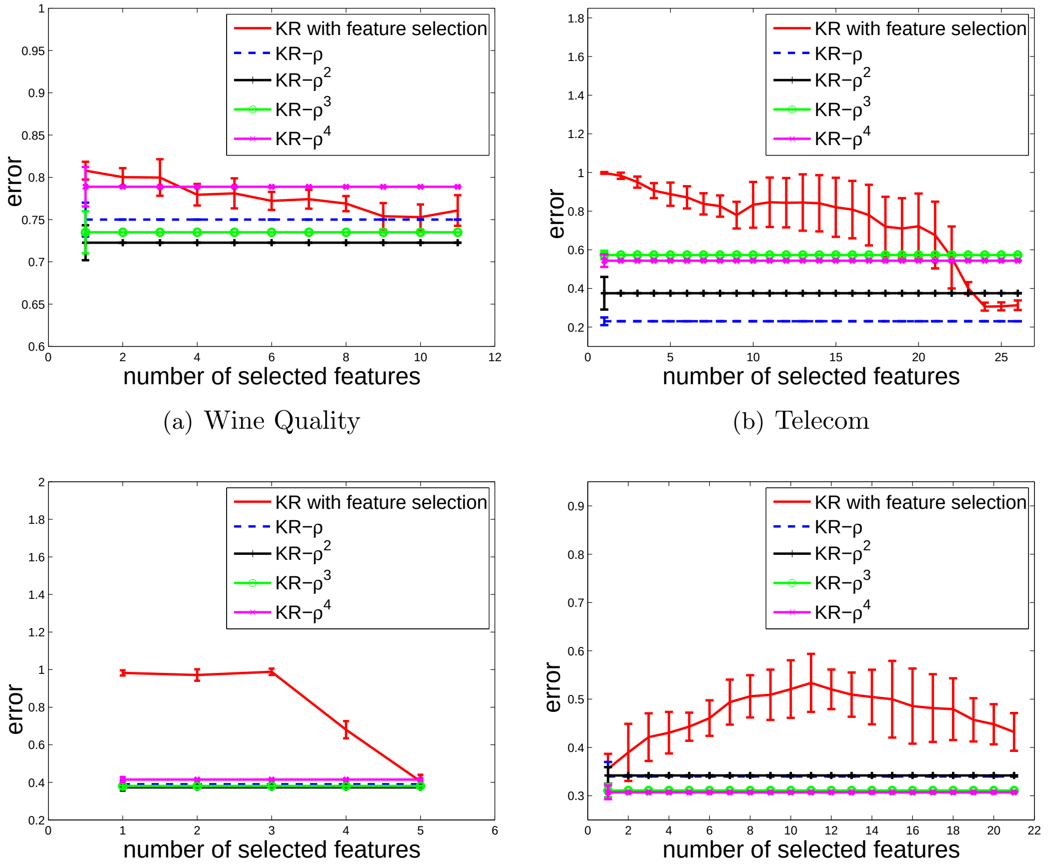
Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Gradients Weights improve Regression and Classification- Supplementary Material

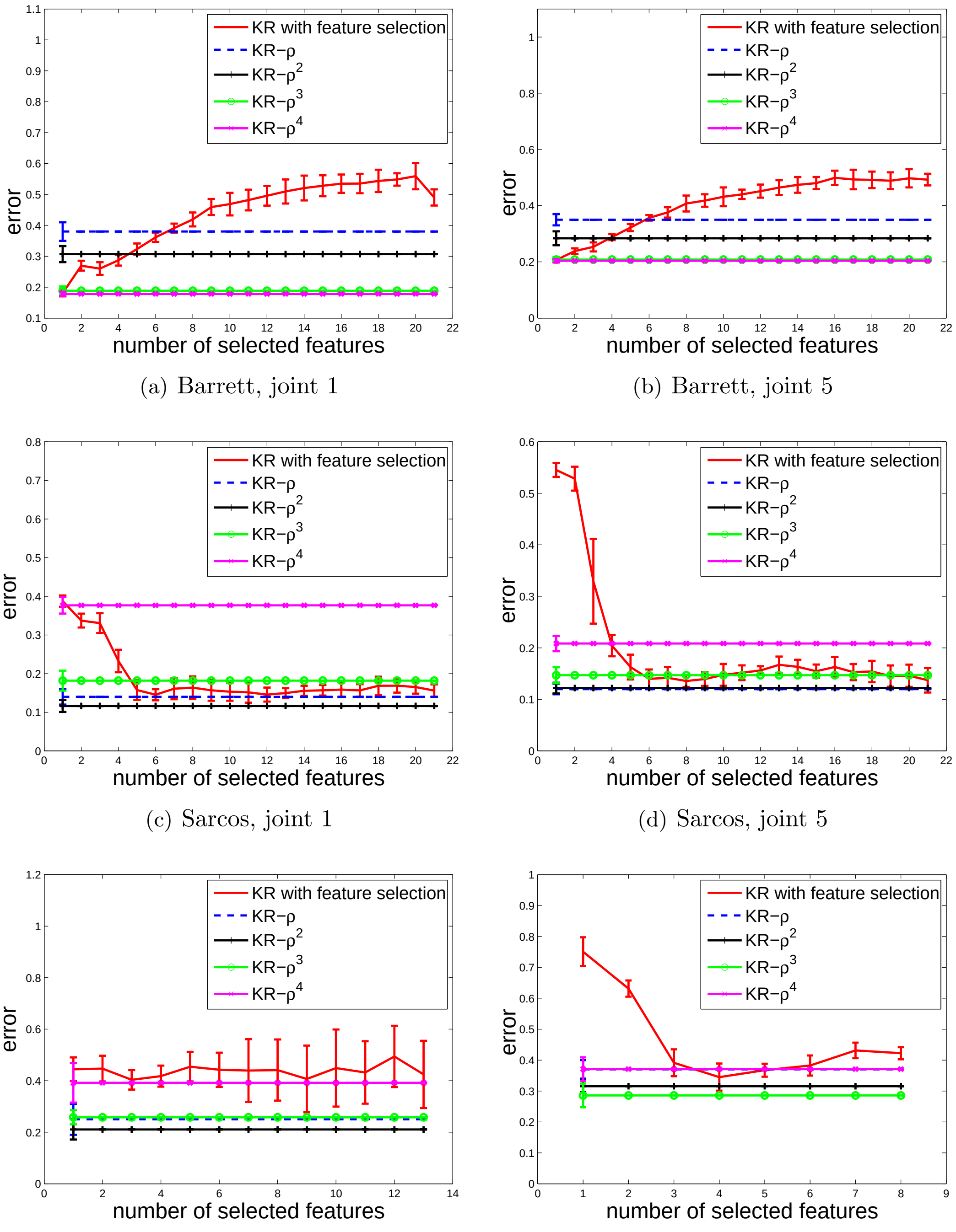
Editor: Somebody



(a) Wine Quality (b) Telecom

(c) Ailerons (d) Parkinson's

Figure 1: Experiments on kernel regression with feature selection vs. gradient weights



(a) Barrett, joint 1

(b) Barrett, joint 5

(c) Sarcos, joint 1

(d) Sarcos, joint 5

(e) Housing

(f) Concrete Strength

Figure 2: Experiments on kernel regression with feature selection vs. gradient weights

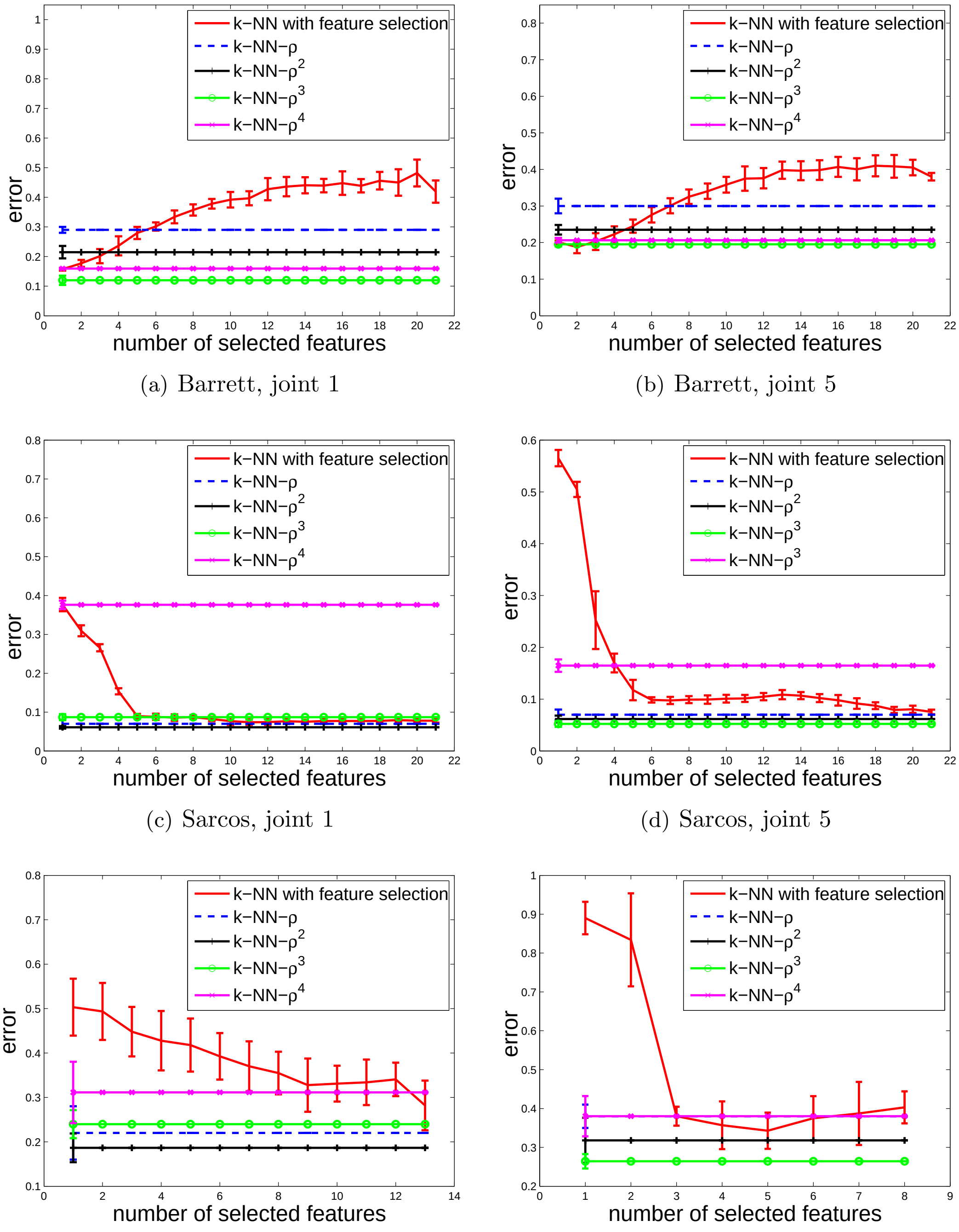
	Barrett joint 1	Barrett joint 5	SARCOS joint 1	SARCOS joint 5	Housing
KR-unnormalized	0.98 ± 0.03	0.90 ± 0.03	0.16 ± 0.02	0.32 ± 0.03	0.73 ± 0.09
KR-r-normalized	0.97 ± 0.01	0.89 ± 0.02	0.19 ± 0.01	0.33 ± 0.04	0.75 ± 0.07
KR-normalized	0.50 ± 0.02	0.50 ± 0.03	0.16 ± 0.02	0.14 ± 0.02	0.37 ± 0.08
KR-normalized-ρ	0.38 ± 0.03	0.35 ± 0.02	0.14 ± 0.02	0.12 ± 0.01	0.25 ± 0.06
KR-normalized-ρ^2	**0.30** ± 0.03	**0.28** ± 0.03	**0.11** ± 0.02	**0.12** ± 0.01	**0.21** ± 0.04
KR-normalized-ρ^3	0.18 ± 0.02	0.20 ± 0.01	0.18 ± 0.03	0.14 ± 0.02	0.25 ± 0.03
KR-normalized-ρ^4	0.17 ± 0.01	0.20 ± 0.01	0.37 ± 0.02	0.20 ± 0.01	0.39 ± 0.08
KR time	0.39 ± 0.02	0.37 ± 0.01	0.28 ± 0.05	0.23 ± 0.03	0.10 ±0.01
KR-ρ time	0.41 ± 0.03	0.38 ± 0.02	0.32 ± 0.05	0.23 ± 0.02	0.11 ±0.01

	Concrete Strength	Wine Quality	Telecom	Ailerons	Parkinson's
KR-unnormalized	0.45 ± 0.03	0.92 ± 0.01	0.23 ± 0.02	0.43 ± 0.02	0.75 ± 0.09
KR-r-normalized	0.43 ± 0.04	0.86 ± 0.02	0.23 ± 0.02	0.45 ± 0.02	0.75 ± 0.06
KR-normalized	0.42 ± 0.05	0.75 ± 0.03	0.30 ± 0.02	0.40 ± 0.02	0.38 ± 0.03
KR-normalized-ρ	0.37 ± 0.03	0.75 ± 0.02	**0.23** ± 0.02	0.39 ± 0.02	**0.34** ± 0.03
KR-normalized-ρ^2	**0.31** ± 0.02	**0.72** ± 0.02	0.37 ± 0.08	**0.37** ± 0.02	**0.34** ± 0.02
KR-normalized-ρ^3	0.28 ± 0.04	0.73 ± 0.03	0.57 ± 0.02	0.38 ± 0.02	0.31 ± 0.02
KR-normalized-ρ^4	0.37 ± 0.04	0.78 ± 0.02	0.54 ± 0.03	0.41 ± 0.01	0.30 ± 0.01
KR time	0.14 ± 0.02	0.19 ± 0.02	0.15±0.01	0.20±0.01	0.30±0.03
KR-ρ time	0.14 ± 0.01	0.19 ± 0.02	0.16±0.01	0.21±0.01	0.30±0.03

	Barrett joint 1	Barrett joint 5	SARCOS joint 1	SARCOS joint 5	Housing
k-NN-unnormalized	0.96 ± 0.01	0.80 ± 0.03	0.11 ± 0.01	0.19 ± 0.01	0.53 ± 0.08
k-NN-r-normalized	0.96 ± 0.01	0.78 ± 0.04	0.12 ± 0.01	0.22 ± 0.02	0.53 ± 0.07
k-NN-normalized	0.41 ± 0.02	0.40 ± 0.02	0.08 ± 0.01	0.08 ± 0.01	0.28 ± 0.09
k-NN-normalized-ρ	0.29 ± 0.01	0.30 ± 0.02	0.07 ± 0.01	0.07 ± 0.01	0.22 ± 0.06
k-NN-normalized-ρ^2	**0.21** ± 0.02	**0.23** ± 0.01	**0.06** ± 0.01	**0.06** ± 0.01	**0.18** ± 0.03
k-NN-normalized-ρ^3	0.11 ± 0.02	0.19 ± 0.01	0.08 ± 0.01	0.05 ± 0.01	0.23 ± 0.03
k-NN-normalized-ρ^4	0.15 ± 0.01	0.20 ± 0.01	0.37 ± 0.01	0.16 ± 0.01	0.31 ± 0.07
k-NN time	0.21 ± 0.04	0.16 ± 0.03	0.13 ± 0.01	0.13 ± 0.01	0.08 ±0.01
k-NN-ρ time	0.13 ± 0.04	0.16 ± 0.03	0.14 ± 0.01	0.13 ± 0.01	0.08 ±0.01

	Concrete Strength	Wine Quality	Telecom	Ailerons	Parkinson's
k-NN-unnormalized	0.40 ± 0.07	0.88 ± 0.01	0.15 ± 0.02	0.42 ± 0.02	0.63 ± 0.04
k-NN-r-normalized	0.37 ± 0.08	0.85 ± 0.02	0.17 ± 0.02	0.44 ± 0.02	0.58 ± 0.02
k-NN-normalized	0.40 ± 0.04	0.73 ± 0.04	**0.13** ± 0.02	0.37 ± 0.01	0.22 ± 0.01
k-NN-normalized-ρ	0.38 ± 0.03	0.72 ± 0.03	0.17 ± 0.02	**0.34** ± 0.01	**0.20** ± 0.01
k-NN-normalized-ρ^2	**0.31** ± 0.06	**0.70** ± 0.01	0.34 ± 0.05	0.34 ± 0.01	**0.20** ± 0.01
k-NN-normalized-ρ^3	0.26 ± 0.02	0.71 ± 0.01	0.55 ± 0.03	0.36 ± 0.01	0.22 ± 0.01
k-NN-normalized-ρ^4	0.38 ± 0.05	0.78 ± 0.01	0.52 ± 0.02	0.45 ± 0.01	0.25 ± 0.01
k-NN time	0.10 ± 0.01	0.15 ± 0.01	0.16±0.02	0.12±0.01	0.14±0.01
k-NN-ρ time	0.11 ± 0.01	0.15 ± 0.01	0.15±0.01	0.11±0.01	0.15±0.01

Table 1: Normalized mean square prediction errors and average prediction time per point (in milliseconds). The top five tables are for KR vs KR-ρ, KR-ρ^2,KR-ρ^3 and KR-ρ^4, the bottom five for k-NN vs k-NN-ρ, k-NN-ρ^2, k-NN-ρ^3 and k-NN-ρ^4. In the methods labeled as normalized, each feature was divided by its empirical standard deviation in the training data. In the methods labeled as unnormalized, we used the raw data sets. In the methods labeled as r-normalized, each data point (row) was normalized so that its l_2 norm is equal to 1. This latter type of normalization does not seem to have a significant benefit. In fact, the prediction errors using unnormalized and r-normalized data are not significantly different.



(a) Barrett, joint 1

(b) Barrett, joint 5

(c) Sarcos, joint 1

(d) Sarcos, joint 5

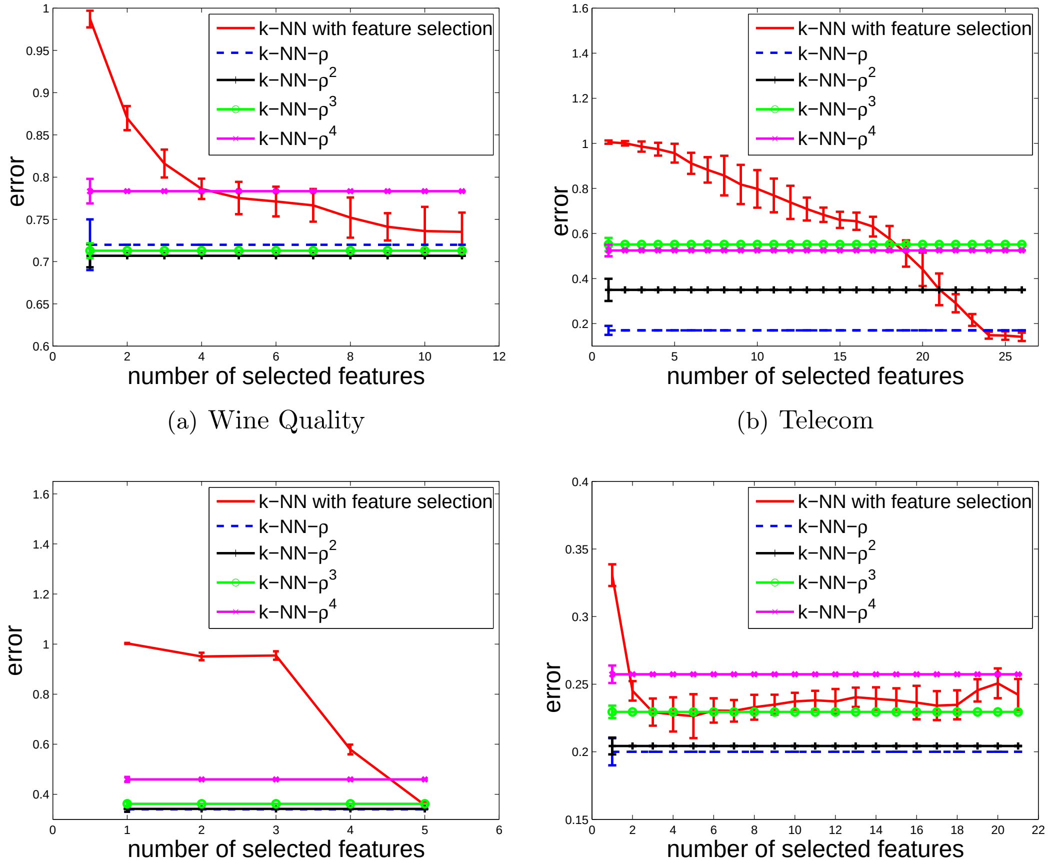
(e) Housing

(f) Concrete Strength

Figure 3: Experiments on k-NN regression with feature selection

	Covertype	IJCNN	MAGIC Gamma	Shuttle	Page Blocks
SVM error	0.25±0.01	0.0576±0.0067	**0.1491**±0.0100	0.0058±0.0027	0.0345±0.0048
SVM-ρ error	**0.24**±0.01	0.0531±0.0059	0.1507±0.0107	**0.0034**±0.0025	0.0342±0.0051
SVM-ρ^2 error	**0.24**±0.01	**0.0521**±0.0065	0.1540±0.0113	**0.0034**±0.0018	**0.0336**±0.0042

Table 2: Classification error rates of a support vector machine suggest that pre-multiplying features by their gradient weight, which corresponds to the ρ^2 metric, also improves performance of that classifier in several cases.



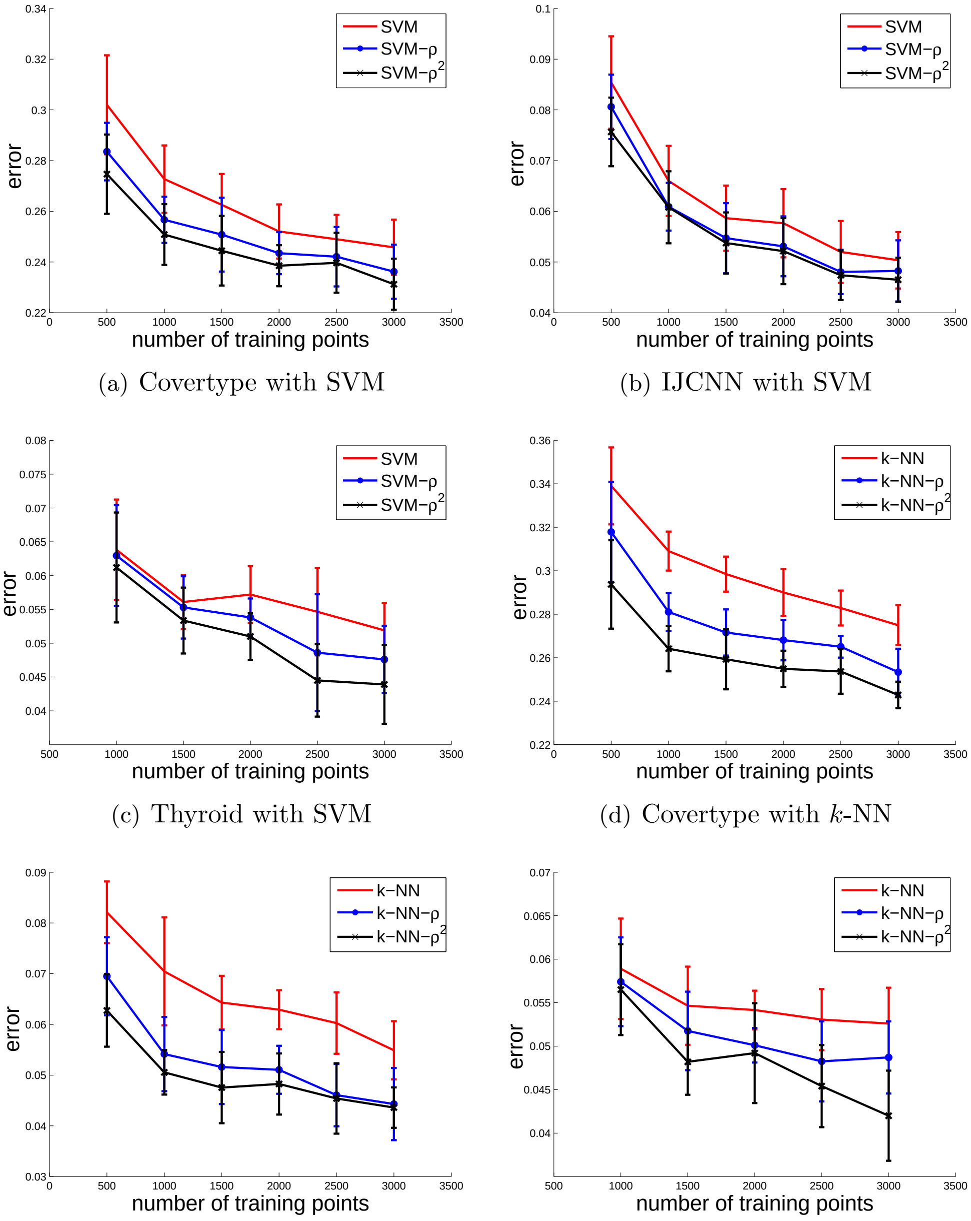
(a) Wine Quality

(b) Telecom

(c) Ailerons

(d) Parkinson's

Figure 4: Experiments on k-NN regression with feature selection



(a) Covertype with SVM

(b) IJCNN with SVM

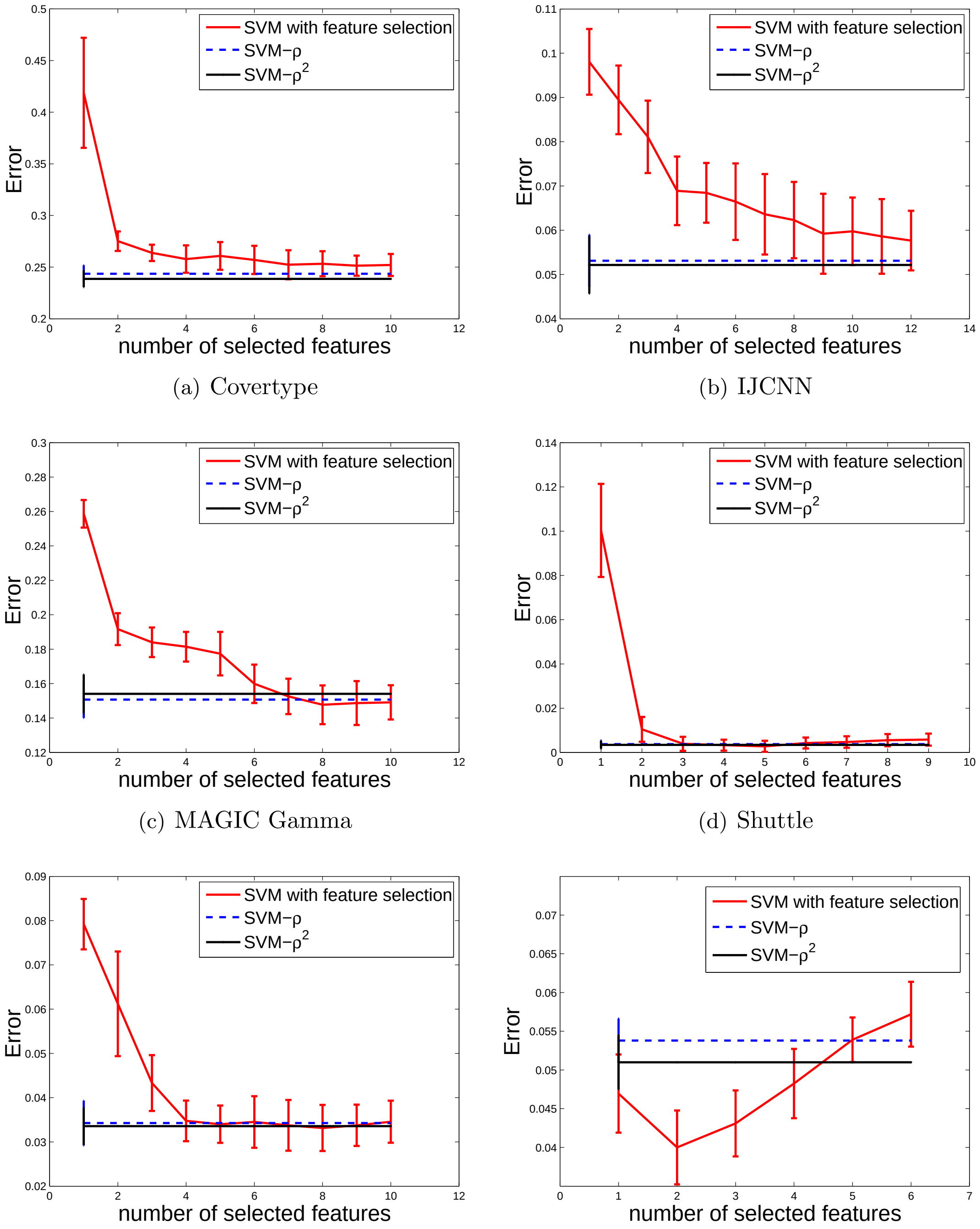
(c) Thyroid with SVM

(d) Covertype with *k*-NN

(e) IJCNN with *k*-NN

(f) Thyroid with *k*-NN

Figure 5: Classification results with gradient weights



(a) Covertype

(b) IJCNN

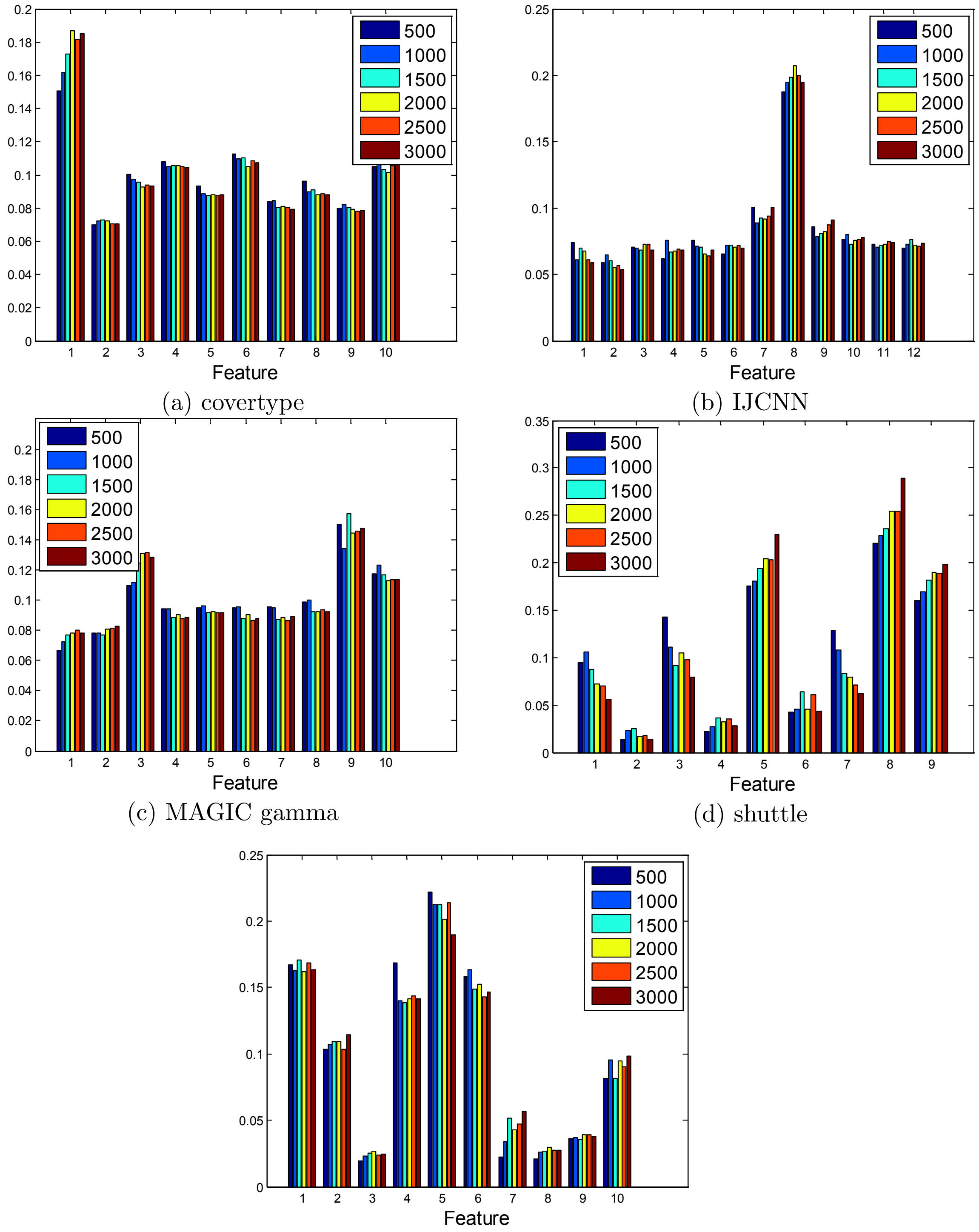
(c) MAGIC Gamma

(d) Shuttle

(e) Page Blocks

(f) Thyroid

Figure 6: Experiments on SVM with feature selection



(a) covertype

(b) IJCNN

(c) MAGIC gamma

(d) shuttle

(e) page blocks

Figure 7: Normalized gradient weights obtained by weighted k-NN with a Gaussian kernel using different training sizes.

Dataset	k-NN	k-NN-ρ
Ailerons	0.3364 ± 0.0087	0.3161 ± 0.0058
Concrete	0.2884 ± 0.0311	**0.2040** ± 0.0234
Housing	0.2897 ± 0.0632	**0.2389** ± 0.0604
Wine	0.6633 ± 0.0119	0.6615 ± 0.0134
Barrett1	0.1051 ± 0.0150	**0.0843** ± 0.0229
Barrett5	0.1095 ± 0.0096	**0.0984** ± 0.0244
Sarcos1	0.1222 ± 0.0074	**0.0769** ± 0.0037
Sarcos5	0.0870 ± 0.0051	0.0779 ± 0.0026
Parkinson	0.3638 ± 0.0443	**0.3181** ± 0.0477
TeleComm	0.0864 ± 0.0094	0.0688 ± 0.0074

Table 3: Regression results, with ten random runs per data set. For each method, the values of k as well as t (the bandwidth used to estimate finite differences for calculating the gradients) were set by two fold cross-validation on the training set.

Dataset	k-NN	k-NN–ρ
Cover Type	0.2279 ± 0.0091	0.2135 ± 0.0064
Gamma	0.1775 ± 0.0070	0.1680 ± 0.0075
Page Blocks	0.0349 ± 0.0042	0.0361 ± 0.0048
Shuttle	0.0037 ± 0.0025	0.0024 ± 0.0016
Musk	0.2279 ± 0.0091	0.2135 ± 0.0064
IJCNN	0.0540 ± 0.0061	0.0459 ± 0.0058
RNA	0.1042 ± 0.0063	0.0673 ± 0.0062

Table 4: Classification results, with ten random runs per data set. For each method, the values of k as well as t (the bandwidth used to estimate finite differences for calculating the gradients) were set by two fold cross-validation on the training set.