Midterm Jeopardy!

Naïve Bayes	K-NN	SVMs	General	General (part 2)	K-means / PCA
10	10	10	10	10	10
20	20	20	20	20	20
30	30	30	30	30	30
40	40	40	40	40	40
50	50	50	50	50	50

Naïve Bayes—10 points

Why is Naïve Bayes called generative and what is its discriminative counter part (in certain settings)?

Naïve Bayes—20 points

[T/F] NB assumes that the features are independent $P(\vec{x}) = \prod_{\alpha} P([x]_{\alpha})$.

Naïve Bayes—30 points

[T,F] If the training dataset is linearly separable, Naïve Bayes will obtain zero training error.

Naïve Bayes—40 points

Give two possible ways to estimate parameters in our models for $p(x_\alpha|y)$ in NB. How do they differ?

Naïve Bayes—50 points

Are Naïve Bayes decision boundaries always linear in the Gaussian case? If not, is there a simple condition that makes them linear?

k-NNs—10 points

[T,F] k-NN is a generative classifier.

k-NNs—20 points

What is the main assumption behind nearest neighbor classification?

k-NNs—30 points

Name one advantage of kNN over logistic regression, and vice versa.

k-NNs—40 points

Why does k-NN still perform well on high dimensional faces and handwritten digits data?

K-NNs—50 points

Assume you have points in 2d within the unit circle and a hyper-plane that dissects the circle.

As you add dimensions with random feature values, how are the pairwise point-distances affected? How are the distances to the hyper-plane affected?

SVMs—10 points

[T,F] Given a data set with two classes, the Perceptron algorithm always finds a separating hyper-plane within finitely many iterations, but it does not necessarily have a large margin.

SVMs—20 points

Why do SVMs maximize the margin?

SVMs—30 points

What is the Perceptron update after a positive input point x is misclassified?

SVMs—40 points

Why is the SVM margin exactly $\frac{1}{||w||_2}$?

SVMs—50 points

Why shouldn't you incorporate the bias as a constant features when working with SVMs?

General—10 points

[T,F] Newton’s method always converges, but sometimes it is slower than Gradient Descent.

General—20 points

Name two assumptions that the Perceptron makes about the data.

General—30 points

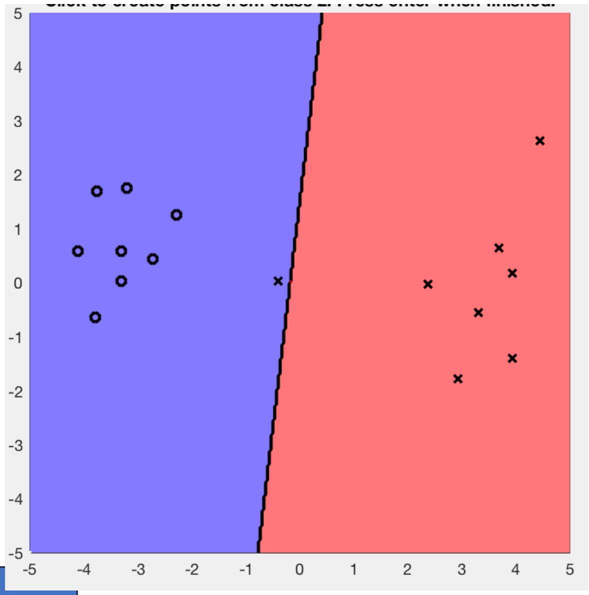
In regression, give one reason to prefer the **squared loss over the absolute loss**, and one reason to prefer the **absolute loss over the squared loss.**

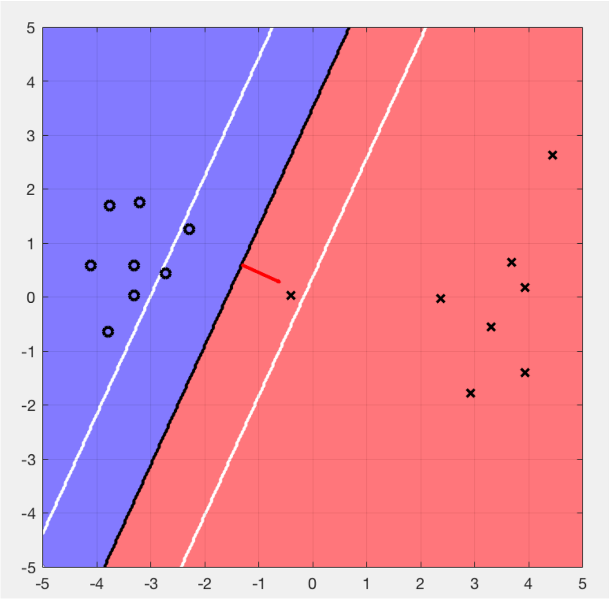
General—40 points

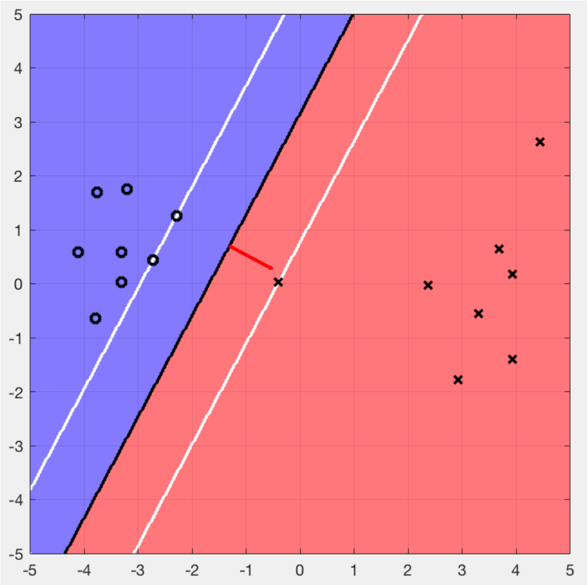
Name the key advantage of Adagrad over the plain Gradient Descent algorithm.

General—50 points

See the SVM decision boundaries below. Which correspond to what value of C.
[C=1, C=0.0001, C=100]







General (part 2)—10 points

Why is your test error typically higher than your validation error?

General (part 2)—20 points

Classify into **discriminative / generative:**

- kNN
- SVM
- Logistic regression
- Perceptron
- Naïve Bayes

General (part 2)—30 points

[T,F] If ML is done right, the data scientist does not have to make any choices—making assumptions is “cheating”.

General (part 2)—40 points

Which algorithm would you expect to result in lower test-time classification error:

kNN or **linear SVM**?

Tasks:

a) Text documents classified by topic?

b) Given patients’ vital signs (15 measurements) predict disease.

c) Handwritten digits classified as 3 vs. 8.

General (part 2)—50 points

For each algorithm name one assumption that it makes on the data:

- **kNN**
- **Naïve Bayes**
- **Logistic regression**
- **SVM**

K-means/PCA—10 points

[T,F] K-means and PCA require labeled data

K-means/PCA—20 points

Give an example of how to choose *k*, the number of clusters, when using k-means.

K-means/PCA—30 points

Why do we remove the mean from data before computing principal components (i.e., equating them to eigenvectors)?

K-means/PCA—40 points

[T,F] K-means algorithm always yields the *optimal* clustering for a given *k.* If it doesn't, why not?

K-means/PCA—50 points

State the two “problems” that the principal components solve given a data set.