Multiple Choice

1. For the Bayes classifier, one has to know the **true** class conditional probabilities.

False: Because we never know the **true** class conditional probability, only estimated one?

2. The Quadratic Discriminant classifier is a generative classifier.

False: Whether we need to estimate some parameters, like mean and covariance. If we need, then it is a generative model; otherwise, it is a discriminant classifier.

Method	Generative	Discriminative	Linear	Non-linear	Parametric	Non-parametric
LDA	✓		✓		✓	
QDA	✓			✓	✓	
Nearest mean	✓		✓		✓	
Parzen	✓			✓		✓
K-nn	✓			✓		✓
Naive Bayes	✓		(✓)	✓	✓	✓
Logistic reg.		✓	✓		✓	
SVM		✓	✓	(✓)	✓	
Decision trees		✓		✓		✓
MLP		✓		✓	✓	

3. The VC dimension is only defined for support vector classifiers.

False: Only some of them are known, like LDC=3. It is hard to compute VC dimension for all classifiers.

4. The logistic classifier assumes that the decision boundary can be modeled by a logistic function.

False: The cost/loss function is logistic formula.

5. k-Means clustering results reproduce exactly when the initialization of the cluster means is non-random.

True: The outputs will be identical if the initialization is fixed with fixed k.

6. In hierarchical clustering, one way to determine the number of clusters is to cut the

dendrogram at its largest jump.

True

7. The Parzen classifier, i.e., the generative classifier in which the underlying classes are modeled by Parzen estimators, is insensitive to feature scaling.

False: They estimate the parameters regardless of the feature scaling. But for discriminative models, they contain some parameters relevant to the feature scaling, like the learning rate.

8. Given the original feature dimensionality is d. If we want to extend the feature space with all possible cross-terms of polynomial degree two, we need **more than** $\frac{1}{2}d(d+1)$ additional dimensions.

False:

$$C_d^2 + d = \frac{1}{2}d(d-1) + d = \frac{1}{2}d(d+1)$$

9. Linear regression is a special form of logistic regression.

False

10. Skip

11. The Bayes error is always larger or equal to the apparent error.

False: always smaller to the apparent error

12. Skip

13. Consider the standard probabilistic form of (unregularized) linear regression based on a Gaussian noise model. Denote the likelihood of this model by p(X|w) with X the observed data. Let q(w) be a pdf that is **uniform** on all values of w for which the L1 norm is smaller than some constant a. The MAP solution, using this prior, is equivalent to L1-regularized least squares regression.

True: **uniform – L1; normal – L2**

(a)

$$p(x,y) = p(y|x)p(x) = \frac{1}{2\sigma\tau\pi} e^{-\left(\frac{(y-(xw+u_0))^2}{2\sigma^2} + \frac{(x-v)^2}{2\tau^2}\right)}$$

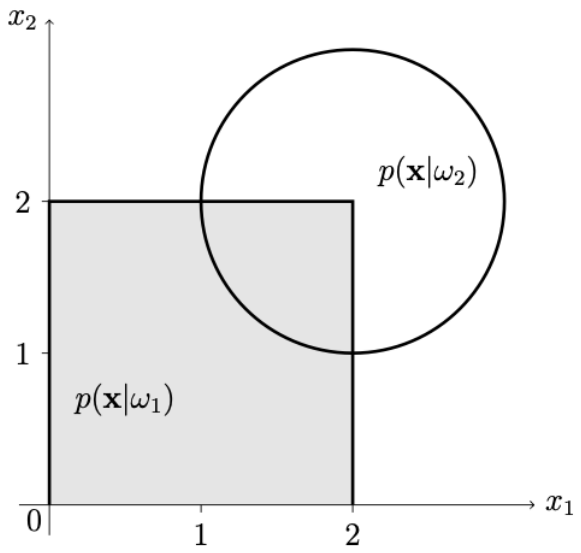
$$p(x,y)p(w) = \frac{1}{\sigma\tau s(2\pi)^{3/2}} e^{-\left(\frac{(y-(xw+u_0))^2}{2\sigma^2} + \frac{(x-v)^2}{2\tau^2} + \frac{w^2}{2s^2}\right)}$$

$$max_w log\ p(x,y)p(w) \simeq min_w\left(\frac{(y-(xw+w_0))^2}{2\sigma^2} + \frac{(x-v)^2}{2\tau^2}\right) + \frac{w^2}{2s^2}\right)$$

$$\rightarrow_{only\ consider\ w\ terms}\ min_w\left(\frac{(y-(xw+w_0))^2}{2\sigma^2} + \frac{w^2}{2s^2}\right)$$

14. Skip

2D Classification



Assume that both classes are equally likely. What is the Bayes error for this problem?

$$0.5 \times \frac{1}{4} \times \frac{\pi \times 1^2}{4} = 0.098$$

Now assume that the prior of class 1 is changed to 0.8. What will be the Bayes error now?

$$0.2 \times \frac{1}{\pi} \times \frac{\pi \times 1^2}{4} = 0.05$$

Assume we fit a logistic classifier: on a very large training set. In which direction will ω point

towards?

$$\omega = [-1, -1]$$

Now we have three classifiers available: (1) the nearest mean classifier, (2) the quadratic classifier and (3) the 1-nearest neighbour classifier. What classifier should you choose for (a) very small training set sizes, and for (b) very large training set sizes?

If we have a small number of training samples, we need a very simple, inflexible and stable classifier – the nearest mean; If we have very many training samples, we can afford a complex, flexible classifier. The most flexible of all given classifiers is the 1NN.

Alternative perceptron classifier

$$\frac{\partial J}{\partial \omega} = \sum_{misclassified\ x_i} \frac{1}{2} \frac{-y_i x_i}{\sqrt{-y_i(w^T x_i + w_0)}} = \frac{1}{2} \frac{[0, -1]}{\sqrt{0.01}} = [0, -5]$$

$$\frac{\partial J}{\partial w_0} = \sum_{misclassified\ x_i} \frac{1}{2} \frac{-y_i}{\sqrt{-y_i(w^T x_i + w_0)}} = \frac{1}{2} \frac{1}{\sqrt{0.01}} = 5$$

$$\omega = \omega - \eta[0, -5] = [1, 0.5]$$

$$w_0 = w - \eta 5 = -0.49$$

PCA

Transpose Axiom:

$$(AB)^T = B^T A^T$$

$$(A^T)^T = A$$

$$(kA)^T = kA^T$$

What is the first principal component of the original data for which we have the covariance matrix C?

The largest variance

Assume we transform all the data by the transformation matrix RR, what does the covariance of the transformed data become?

$$RX(RX)^T = RXX^T R^T = RCR^T$$

PCA Procedures:

Assume we transform all the data by the transformation matrix R, what does the covariance of the transformed data become?

1D Regression

Rather than just fitting a least-squares model, we consider a maximum likelihood solution under an assumed Gaussian noise model. That is, we assume that outputs are obtained as a function f from x plus some fixed-variance, independent Gaussian noise.

If our fit to the 5 data point equals the constant zero function, i.e., f(x)=0, what then is the maximum likelihood estimate for the variance of the Gaussian noise?

The variance equal 1-precision and is simply estimated by the average squared loss achieved on the training data, i.e., (0 + 0 + 1 + 0 + 1)/5 = 2/5.

https://towardsdatascience.com/maximum-likelihood-estimation-explained-normal-distribution-6207b322e47f

$$\varepsilon \sim N(0, \sigma^2)$$

$$w^T x + \epsilon = y \sim N(w^T x, \sigma^2)$$

Maximum Likelihood Estimator:

$$\mu = \frac{1}{n}\sum y_i$$

$$\sigma^2 = \frac{1}{n}\sum (y_i - \mu)^2$$

Curves

When the number of features increases: Each of the features contributes a bit to the discrimination between the classes. So if we would know the distributions perfectly (which a Bayes classifier does), the class overlap would decrease and decrease, and the Bayes error decreases as well.

When the number of features increases: The classifier is trained on some training data,

which is finite. So at a certain moment it will suffer from the curse of dimensionality, and the performance will deteriorate. The true error will first go down (more useful information) and later goes up (overfitting in a too large feature space).

When the number of features decrease: Feature reduction may be tried to combat the curse of dimensionality, but when you push it too far (you increase the number of features further and further), it will anyway suffer from the curse. Fundamentally nothing has changed; first the true error goes down, but at a certain moment it will increase again.