CSC 311: Introduction to Machine Learning

Final Exam Review

University of Toronto

Ensemble Methods

Question: Recall that in bagging, we compute an average of the predictions $y_{\text{avg}} = \frac{1}{m}\sum_{i=1}^{m} y_i$. Recall that these predictions are not fully independent, i.e., they are correlated because their training sets come from the same underlying dataset. Suppose $\text{Var}[y_i] = \sigma^2$ and the correlation between y_i and y_j is ρ for $i \neq j$. Calculate the variance $\text{Var}[y_{\text{avg}}]$.

Ensemble Methods

First, note that

$$\mathrm{Var}(y_{\mathrm{avg}}) = \mathrm{Var}(\frac{1}{m}\sum_{i=1}^{m} y_i)$$
$$\frac{1}{m^2}\mathrm{Var}(\sum_{i=1}^{m} y_i) = \frac{1}{m^2}\mathrm{Cov}(\sum_{i=1}^{m} y_i, \sum_{i=1}^{m} y_i)$$

Now, since Covariance is a linear operation, we'll have

$$\begin{aligned}\mathrm{Cov}(\sum_{i=1}^{m} y_i, \sum_{j=1}^{m} y_j) = \sum_{i=1}^{m}\mathrm{Cov}(y_i, \sum_{j=1}^{m} y_j) &= \sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{Cov}(y_i, y_j) \\ &= \sum_{i=1}^{m}\mathrm{Var}(y_i) + \sum_{i\neq j}\mathrm{Cov}(y_i, y_j) \\ &= m\sigma^2 + m(m-1)\rho\sigma^2\end{aligned}$$

Therefore,

$$\mathrm{Var}(y_{\mathrm{avg}}) = \frac{1}{m^2}\left[m\sigma^2 + m(m-1)\rho\sigma^2\right] = \frac{1}{m}\sigma^2 + \frac{m-1}{m}\rho\sigma^2$$

Question: Suppose your classifier achieves poor accuracy on both the training and test sets. Does bagging improve the performance? Justify your answer.

Question: Suppose your classifier achieves poor accuracy on both the training and test sets. Does bagging improve the performance? Justify your answer.

- The model is underfitting, and has a high bias.
- Bagging reduces variance but does not change the bias.
- Therefore, We wouldn't get a performance boost using bagging.

Probabilistic Models: Naive Bayes

Question: True or False: Naive Bayes assumes that all features are independent.

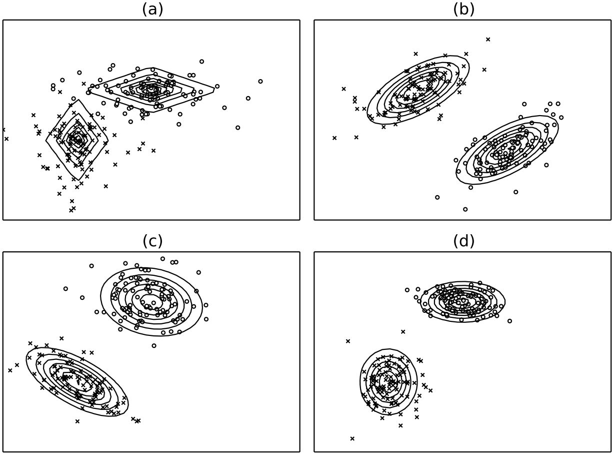
Probabilistic Models: Naive Bayes

Question: True or False: Naive Bayes assumes that all features are independent.

False. Naive Bayes assumes that the input features x_i are **conditionally independent** give the class c:

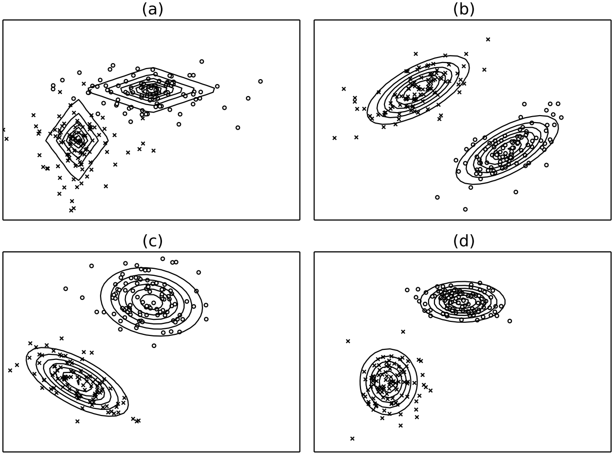
$$p(c, x_1, \ldots, x_D) = p(c)p(x_1|c) \cdots p(x_D|c)$$

Question: Which of the following diagrams could be a visualization of a Naive Bayes classifier? Select all that apply.



Probabilistic Models: Naive Bayes

Question: Which of the following diagrams could be a visualization of a Naive Bayes classifier? Select all that apply.



Answer: A, D

Principal Component Analysis (PCA)

Recall that the PCA code vector for a data point $\mathbf{x}$ is given by $\mathbf{z} = \mathbf{U}^\top(\mathbf{x} - \hat{\boldsymbol{\mu}})$. Show that the entries of $\mathbf{z}$ are uncorrelated.

Principal Component Analysis (PCA)

Recall that the PCA code vector for a data point $\mathbf{x}$ is given by $\mathbf{z} = \mathbf{U}^\top(\mathbf{x} - \hat{\boldsymbol{\mu}})$. Show that the entries of $\mathbf{z}$ are uncorrelated.

Answer:

$$\begin{aligned}
\mathrm{Cov}(\mathbf{z}) &= \mathbb{E}\left[(\mathbf{z} - \mathbb{E}[\mathbf{z}])(\mathbf{z} - \mathbb{E}[\mathbf{z}])^\top\right] \\
&= \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\right] \\
&= \mathbf{U}^\top \mathbb{E}\left[(\mathbf{x} - \hat{\boldsymbol{\mu}})(\mathbf{x} - \hat{\boldsymbol{\mu}})^\top\right] \mathbf{U} \\
&= \mathbf{U}^\top \hat{\boldsymbol{\Sigma}} \mathbf{U} \\
&= \mathbf{U}^\top \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^\top \mathbf{U} \\
&= \begin{pmatrix} \mathbf{I} & 0 \end{pmatrix} \boldsymbol{\Lambda} \begin{pmatrix} \mathbf{I} \\ 0 \end{pmatrix}
\end{aligned}$$

Which is the top $K \times K$ block of $\boldsymbol{\Lambda}$. Matrix $\boldsymbol{\Lambda}$ is diagonal $\implies$ Uncorrelated features

Principal Component Analysis (PCA)

Consider the following data matrix, representing four samples $X_i \in \mathbb{R}^2$:

$$\mathbf{X} = \begin{pmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{pmatrix}$$

1. Compute the unit-length principal component directions of $\mathbf{X}$, and state which one the PCA algorithm would choose if you request just one principal component.
2. Find the best (min reconstruction error) projection of $\mathbf{X}$ into a 1-dimensional subspace with the origin of zero.

Principal Component Analysis (PCA)

1. We first center the data matrix, yielding

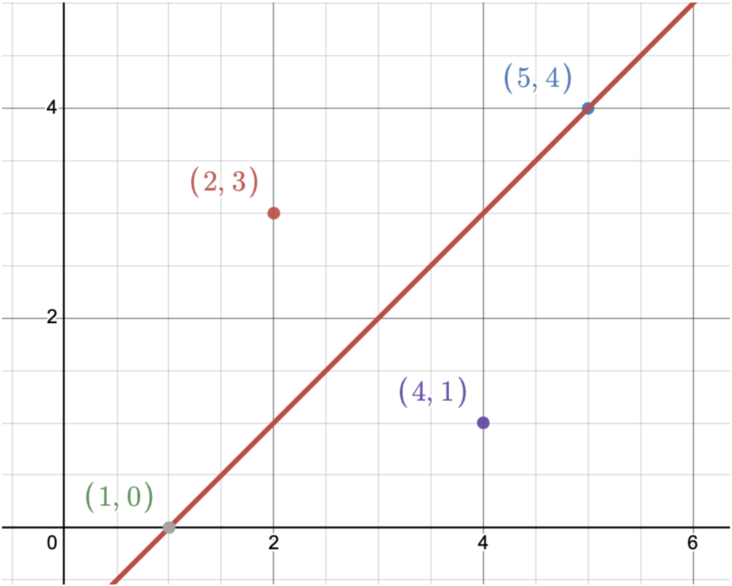
$$\hat{X} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{pmatrix}$$

We then calculate the empirical covariance

$$\frac{1}{4}\hat{X}^\top \hat{X} = \frac{1}{4}\begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$$

The eigenvectors are $\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}^\top$ with eigenvalue 16 and $\begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}^\top$ with eigenvalue 4. The former eigenvector is chosen.

Principal Component Analysis (PCA)



Principal Component Analysis (PCA)

2. Recall that we showed the following equivalence in the lecture

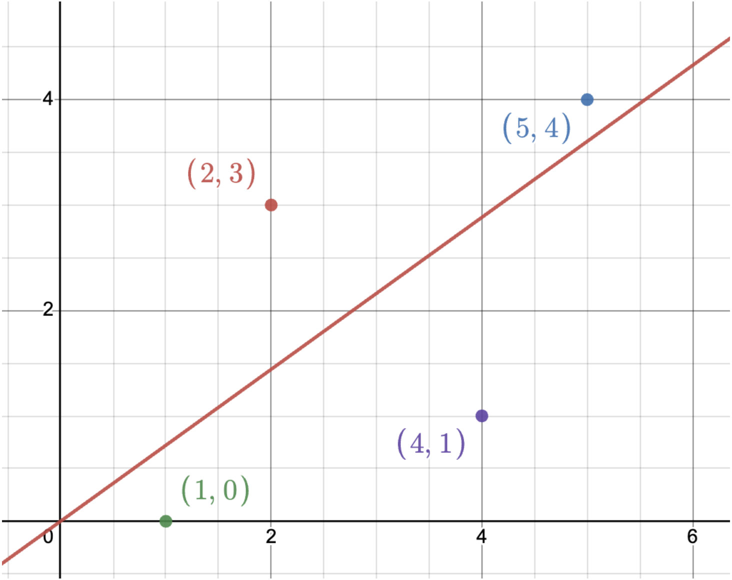
$$\min_{\mathbf{U}} \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|^2 \equiv \max_{\mathbf{U}} \frac{1}{N} \sum_{i=1}^{N} \|\hat{\mathbf{x}}^{(i)} - \hat{\mu}\|^2$$

However, in the proof of the equivalence, we didn't use any property of $\hat{\mu}$ being the center of the data. Therefore, we can consider $\hat{\mu} = 0$ for this problem. The only difference is that we won't center the data $\mathbf{X}$:

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 46 & 30 \\ 30 & 26 \end{pmatrix}$$

The eigenvectors corresponding to the largest eigenvalue is $\begin{pmatrix} \frac{1+\sqrt{10}}{3} & 1 \end{pmatrix}^\top$.

Principal Component Analysis (PCA)



Probabilistic Models

The Laplace distribution, parameterized by μ and β, is defined as follows:

$$\text{Laplace}(w; \mu, \beta) = \frac{1}{2\beta} \exp\left(-\frac{|w - \mu|}{\beta}\right)$$

We have a labeled training set $\mathcal{D} = \{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^{N}$ and the goal is to predict target t from covariates x. We assume a linear Gaussian model for the target variable, i.e.,

$$t|\mathbf{w} \sim \mathcal{N}(t; \mathbf{w}^\top \mathbf{x}, \sigma^2)$$

We assume the following prior over the weights $\mathbf{w}$:

$$w_j \sim \text{Laplace}(0, \beta)$$

The Gaussian PDF is:

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Probabilistic Models

1. Give the cost function you would minimize to find the MAP estimate of $\mathbf{w}$.

To find the MAP estimation, we first write down the posterior distribution

$$\begin{aligned}
\text{posterior}(\mathbf{w}|\mathcal{D}) &\propto P(\mathcal{D}|\mathbf{w}) \cdot \text{prior}(\mathbf{w}) \\
&\propto \prod_{i=1}^{N} P(t^{(i)}|x^{(i)};\mathbf{w}) \cdot \prod_{j} \exp\left(-\frac{|w_j|}{\beta}\right) \\
&\propto \prod_{i=1}^{N} \exp\left(-\frac{(t^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)})^2}{2\sigma^2}\right) \cdot \prod_{j} \exp\left(-\frac{|w_j|}{\beta}\right)
\end{aligned}$$

The MAP estimator is as follows:

$$\begin{aligned}
\mathbf{w}_{\text{MAP}} &= \underset{\mathbf{w}}{\operatorname{argmax}} \; \log \text{posterior}(\mathbf{w}|\mathcal{D}) \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{\beta} \sum_{j} |w_j| + \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(t^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)}\right)^2
\end{aligned}$$

Probabilistic Models: Naïve Bayes

Question:

- Consider the following problem, in which we have two classes: $\{\text{Tainted}, \text{Clean}\}$, and three covariate features: (a_1, a_2, a_3).
- These attributes are also binary variables: $a_1 \in \{\text{on}, \text{off}\}$, $a_2 \in \{\text{blue}, \text{red}\}$, $a_3 \in \{\text{light}, \text{heavy}\}$.
- We are given a training set as follows:
  1. Tainted: (on, blue, light) (off, red, light) (on, red, heavy)
  2. Clean: (off, red, heavy) (off, blue, light) (on, blue, heavy)

(A) Manually construct Naïve Bayes Classifier based on the above training data. Compute the following probability tables:

a The class prior probability

b The class conditional probabilities of each attribute.

Probabilistic Models: Naïve Bayes

(a) Class prior probability:

- $p(c = \text{Tainted}) = 3/6 = 1/2$,
- $p(c = \text{Clean}) = 1/2$

Probabilistic Models: Naïve Bayes

(a) Class prior probability:

- $p(c = \text{Tainted}) = 3/6 = 1/2$,
- $p(c = \text{Clean}) = 1/2$

(b) The class conditional distributions:

- $p(a_1 = \text{on}|c = \text{Tainted}) = 2/3$, $p(a_1 = \text{off}|c = \text{Tainted}) = 1/3$

Probabilistic Models: Naïve Bayes

(a) Class prior probability:

- $p(c = \text{Tainted}) = 3/6 = 1/2,$
- $p(c = \text{Clean}) = 1/2$

(b) The class conditional distributions:

- $p(a_1 = \text{on}|c = \text{Tainted}) = 2/3$, $p(a_1 = \text{off}|c = \text{Tainted}) = 1/3$
- $p(a_2 = \text{blue}|c = \text{Tainted}) = 1/3$, $p(a_2 = \text{red}|c = \text{Tainted}) = 2/3$
- $p(a_3 = \text{light}|c = \text{Tainted}) = 2/3$, $p(a_3 = \text{heavy}|c = \text{Tainted}) = 1/3$
- $p(a_1 = \text{on}|c = \text{Clean}) = 1/3$, $p(a_1 = \text{off}|c = \text{Clean}) = 2/3$
- $p(a_2 = \text{blue}|c = \text{Clean}) = 2/3$, $p(a_2 = \text{red}|c = \text{Clean}) = 1/3$
- $p(a_3 = \text{light}|c = \text{Clean}) = 1/3$, $p(a_3 = \text{heavy}|c = \text{Clean}) = 2/3$

(B) Classify a new example $(\text{on}, \text{red}, \text{light})$ using the classifier you built above. You need to compute the posterior probability (up to a constant) of class given this example.

(B) Classify a new example $(\text{on}, \text{red}, \text{light})$ using the classifier you built above. You need to compute the posterior probability (up to a constant) of class given this example.

Answer: To classify $\mathbf{x} = (\text{on}, \text{red}, \text{light})$, we have:

$$p(c|\mathbf{x}) = \frac{p(c)p(x|c)}{p(c = \text{Tainted})p(x|c = \text{Tainted}) + p(c = \text{Clean})p(x|c = \text{Clean})}$$

Computing each term:

$$\begin{aligned} p(c = T)p(x|c = T) &= p(c = T)p(a_1 = on|c = T)p(a_2 = \text{red}|c = T) \\ &\quad p(a_3 = \text{light}|c = T) \\ &= \frac{1}{2} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \\ &= \frac{8}{54} \end{aligned}$$

Probabilistic Models: Naïve Bayes

(B) Classify a new example $(\text{on}, \text{red}, \text{light})$ using the classi
er you built above. You need to compute the posterior probability (up to a constant) of class given this example.

Answer: Similarly,

$$p(c = \text{Clean})p(x|c = \text{Clean}) = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{54}$$

Therefore, $p(c = \text{Tainted}|\mathbf{x}) = 8/9$ and $p(c = \text{Clean}|\mathbf{x}) = 1/9$. According to Naïve Bayes classifier this example should be classified as **Tainted**.