Statistical learning - HW 1

Ex 2.2

a) The prosecutor claims that

$$p(Btype = Crime \mid Inocent) = 0{,}01$$

and confuses this with

$$p(Inocent \mid Btype = Crime) = 0{,}01$$

and because this probability is low than he concludes that the probability of being guilty knowing that possesses the guilty blood type is high.

But we know by Bayes theorem that

$$p(Btype = Crime \mid Inocent) \neq P(Inocent \mid Btype = Crime)$$

so he is wrong.

b) The defender claims that the probability of being guilty knowing that the defendent has the guilty blood type is only 1/8000 and that it is not relevant, though it is relevant! The introduction of the blood type knoledge was able to shring the guilty space from 800000 to 8000.

Ex. 2.7

Lets consider the example of the toss of a balanced coin twice, therefore the probability of heads (H) and tails (T) is 0,5.

Lets define the following events:

A - having the first toss to be heads, $\{HH, HT\}$

B - having the second toss to be heads, $\{HH, TH\}$

C - having both tosses the same, $\{HH, TT\}$

The probability of each event is:

$$P(A) = P(B) = P(C) = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2}$$

The probability of $A \cap B$, $B \cap C$ and $A \cap C$ are:

$$P(A \cap B) = \frac{1}{2}\cdot\frac{1}{2} = P(A)\cdot P(B) = \frac{1}{4}$$

$$P(B \cap C) = \frac{1}{2}\cdot\frac{1}{2} = P(B)\cdot P(C) = \frac{1}{4}$$

$$P(A \cap C) = \frac{1}{2}\cdot\frac{1}{2} = P(A)\cdot P(C) = \frac{1}{2}$$

This means the events are pair wise independent.

Now lets check for mutual independence:

$$P(A \cap B \cap C) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \neq P(A)\cdot P(B)\cdot P(C) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}$$

So they are not mutualy independent though they are pair wise!

Ex 4.2

a)

$$P(Y < a) = E\{P(Y<a)\} = E\{P(WX < a)\} = \sum_{\bar{w}\in W} P(\bar{w}X < a)\cdot P_W(\bar{w})$$

$$= P(-X < a)\cdot P(-1) + P(X < a)\cdot P_W(1) =$$

$$= \frac{1}{2}\cdot P(X<a) + \frac{1}{2} P(X<a) = P(X<a), \forall a \in \mathbb{R}$$

Therefore both follow the same distribution.

b)

$$\operatorname{cov}[X,Y] = E[XY] - \underbrace{E[X]}_{=0} E[Y] = E[E[XY|W]] =$$

$$= E[E[X^2 W|W]] = E\left[\sum_{\bar{w}\in W} X^2\bar{w}\cdot P_W(\bar{w})\right] =$$

$$= E\left[-X^2\cdot 0{,}5 + X^2\cdot 0{,}5\right] = E[0] = 0$$

Ex 5.3

a)

$$\delta(x) = \arg\min_{a \in A} \rho(a|x) = \arg\min_{a \in A} \sum_{y \in \{1,\dots,c+1\}} L(y,a) \cdot p(y|x) =$$

$$= \arg\min_{a \in A} L(1,a) \cdot p(1|x) + \dots + L(j,a) \cdot p(j|x) + \dots + L(c,a) \cdot p(c|x)$$

• We choose $a = \alpha_j$ if $\rho(\alpha_j|x) \leq \rho(\alpha_i|x)$ $\forall i \in \{1,\dots,c\} / \{j\}$

$\rho(\alpha_j|x) \leq \rho(\alpha_i|x) \Leftrightarrow$

$$\Leftrightarrow \underbrace{L(1,\alpha_j)}_{=\lambda_s} \cdot p(1|x) + \dots + \underbrace{L(j,\alpha_j)}_{=0} p(j|x) + \dots + \underbrace{L(i,\alpha_j)}_{=\lambda_s} \cdot p(i|x) + \dots + \underbrace{L(c,\alpha_j)}_{=\lambda_s} \cdot p(c|x) \leq \underbrace{L(1,\alpha_i)}_{=\lambda_s} \cdot p(1|x) + \dots + \underbrace{L(j,\alpha_i)}_{=\lambda_s} \cdot p(j|x) + \dots + \underbrace{L(i,\alpha_i)}_{=0} \cdot p(i|x) + \dots + \underbrace{L(c,\alpha_i)}_{=\lambda_s} \cdot p(c|x) \Leftrightarrow$$

$$\Leftrightarrow \lambda_s \cdot p(i|x) \leq \lambda_s \cdot p(j|x) \Leftrightarrow p(j|x) \geq p(i|x)$$

• And also if $\rho(\alpha_j|x) \leq \rho(\alpha_{c+1}|x)$

$\rho(\alpha_j|x) \leq \rho(\alpha_{c+1}|x) \Leftrightarrow$

$$\Leftrightarrow \underbrace{L(1,\alpha_j)}_{=\lambda_s} \cdot p(1|x) + \dots + \underbrace{L(j,\alpha_j)}_{=0} \cdot p(j|x) + \dots + \underbrace{L(i,\alpha_j)}_{=\lambda_s} \cdot p(i|x) + \dots + \underbrace{L(c+1,\alpha_j)}_{=\lambda_s} \cdot p(c+1|x) \leq \lambda_r \Leftrightarrow$$

$$\Leftrightarrow \lambda_s \cdot \sum_{\substack{i=1 \\ i \neq j}}^{c} p(i|x) \leq \lambda_r \Leftrightarrow \lambda_s (1 - p(j|x)) \leq \lambda_r \Leftrightarrow$$

$$\Leftrightarrow p(j|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

b) As λ_r/λ_s increases the risk of rejection grows porportionaly to the risk of missclassification.

when $\lambda_r/\lambda_s = 0$, i.e., $\lambda_r = 0$ there is no risk in rejecting so the hypotesis is allways rejected ($P(Y=j|x)$ can't be bigger than 1)

when $\lambda_r/\lambda_s = 1$, then $1 - \frac{\lambda_r}{\lambda_s} = 0$ and no hypotesis is ever rejected

Ex. 7.9

a) Multinomial distribution

$$p(x, y; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left(-\frac{1}{2}(v-\mu)^T \Sigma^{-1}(v-\mu)\right)$$

where:

$$v = \begin{bmatrix} y \\ x \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$$

Maximum Likelihood

of μ and Σ

- maximize distribution of N points

$$p(x_1, y_1, \ldots, x_N, y_N; \mu, \Sigma) = \prod_{i=1}^{N} p(x_i, y_i; \mu, \Sigma) =$$

$$= \left(\frac{1}{\sqrt{(2\pi)^n |\Sigma|}}\right)^N \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(v_i-\mu)^T \Sigma^{-1}(v_i-\mu)\right)$$

- $L = \log(\cdot)$

$$L = -\frac{N}{2}\left[\log(2\pi)^n + \log|\Sigma|\right] - \frac{1}{2}\sum_{i=1}^{N}(v_i-\mu)^T \Sigma^{-1}(v_i-\mu)$$

- $\hat{\mu}$ - first order condition

$$\frac{\partial L}{\partial \hat{\mu}} = 0 \Leftrightarrow \sum_{i=1}^{N}(v_i-\hat{\mu})^T \Sigma^{-1} = 0 \Leftrightarrow \sum_{i=1}^{N}(v_i) - N \cdot \hat{\mu} = 0 \Leftrightarrow$$

$$\Leftrightarrow \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} v_i$$

$$\Rightarrow \hat{\mu} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} y_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i \end{bmatrix}$$

- $\hat{\Sigma}$ - first order conditions

using: $\frac{\partial}{\partial A} x^T A x = \frac{\partial}{\partial A} tr(x^T A x)$

$$= \frac{\partial}{\partial A} tr(x x^T A)$$

$$= (x x^T)^T = x x^T$$

$$\frac{\partial}{\partial A} \log |A| = A^{-1}$$

$$\frac{\partial L}{\partial \hat{\Sigma}^{-1}} = 0 \iff \frac{N}{2} \cdot \hat{\Sigma} - \frac{1}{2} \sum_{i=1}^{N} (v_i - \mu)(v_i - \mu)^T = 0 \iff$$

$$\iff \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (v_i - \mu)(v_i - \mu)^T$$

$$\iff \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} y_i - \mu_y \\ x_i - \mu_x \end{bmatrix} \begin{bmatrix} y_i - \mu_y \\ x_i - \mu_x \end{bmatrix}^T$$

$$\iff \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} (y_i - \mu_y)(y_i - \mu_y)^T & (y_i - \mu_y)(x_i - \mu_x)^T \\ (x_i - \mu_x)(y_i - \mu_y)^T & (x_i - \mu_x)(x_i - \mu_x)^T \end{bmatrix}$$

$\hat{\Sigma}_{xx} = \frac{1}{N} X_c^T \cdot X_c$ $\quad \Sigma_{YY} = \frac{1}{N} Y_c^T Y_c$

$\hat{\Sigma}_{xy} = \frac{1}{N} X_c^T Y_c$ $\quad \hat{\Sigma}_{yx} = \frac{1}{N} Y_c^T X_c$

$\hat{\mu}_x = \frac{1}{N} X^T \cdot \mathbb{1}$

$\hat{\mu}_Y = \frac{1}{N} \cdot Y^T \cdot \mathbb{1}$

$P(Y|X) = \mathcal{N}(Y | \mu_{Y|X}, \Sigma_{Y|X})$ $\quad \mu_{Y|X} = E\{Y|X\}$

$$E\{Y|X\} = \hat{\mu}_Y + \Sigma_{YX} \Sigma_{XX}^{-1} (x - \hat{\mu}_x)$$

$$= \Sigma_{YX} \Sigma_{XX}^{-1} x + \hat{\mu}_Y - \Sigma_{YX} \Sigma_{XX}^{-1} \hat{\mu}_x$$

$$= \frac{1}{\cancel{N}} \cdot Y_c^T X_c \cdot \cancel{N} (X_c^T \cdot X_c)^{-1} x + \hat{\mu}_Y - \frac{1}{\cancel{N}} \cdot Y_c^T X_c \cancel{N} (X_c^T X_c)^{-1} \hat{\mu}_x + \hat{\mu}_Y$$

$$= \underbrace{Y_c^T X_c (X_c^T X_c)^{-1}}_{W^T} \cdot x + \underbrace{\hat{\mu}_Y - \underbrace{Y_c^T X_c (X_c^T X_c)^{-1}}_{W^T} \cdot \hat{\mu}_x}_{W_0^T}$$

5)

My answer will evaluate the advantages and disadvantages of generative and discriminative aproaches generally and not only to linear regression.

- Easy to fit?

It is usually easier to fit generative classifiers.

Ex: Naive Bayes model can be fitted by only counting while logistic regression requires solving a complex optimization problem.

- Fit classes separatly?

In generative classifiers, the parameters of each class are estimated independently, so there is no need to retrain the model when adding more classes.

In discriminative models all the parameters interact, so the whole model must be retrained when inserting a new class

- Handle missing features easily?

In the discriminative models there is no solution to deal with missing features while in generative ones there is.

- Can handle unlabeled training data?

Semi-supervised learning is much easier to do with generative models than discriminative.

- Symmetric in inputs and outputs?

A generative model can be run "backwards", and infer probable inputs given the output by computing $p(X|y)$. Not possible with the discriminative model.

- <u>Can handle feature preprocessing?</u>

A big advantage of discriminative methods is that they allow the preprocess of the data in arbitrary ways.

It is hard to define a generative model on pre-processed data, since the new features are correlated in complex ways.

- <u>Well calibrated probabilities?</u>

Some generative models, such as the naive Bayes, make strong assumption wich are often not valid. This can lead to extreme posterior class probabilities.

Discriminative models are usually better calibrated in terms of their probability estimates