Problem 5

In this problem, we will do regression for data that are generated from a Gaussian Mixture Model. Let $X \in \mathbb{R}$ be the random variable for the features and $Y \in \mathbb{R}$ be the random variable for the output. We assume X is generated from a mixture of m Gaussian distributions, and Y is linearly correlated to X with some random noise. The generation process can be described as follows:

1. Sample a random variable T from a multinomial distribution on $\{1, 2, \ldots, m\}$, where $P(T = t) = p_t$.

2. Sample X from the tth Gaussian distribution, with mean μ_t and variance σ_t^2.

3. Given X from the tth Gaussian distribution, let $Y = w_t X + \epsilon$, where w_t is a fixed parameter for the tth Gaussian and ϵ is from an independent Gaussian with 0 mean and variance of 1.

5. (1)

1/1 point (graded)
Which of the following is the correct probability density of X?

- $\sum_{t=1}^{m} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(X-\mu_t)^2}{2\sigma_t^2}\right)$
- $\sum_{t=1}^{m} \frac{p_t}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(X-\mu_t)^2}{2\sigma_t^2}\right)$
- $\prod_{t=1}^{m} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(X-\mu_t)^2}{2\sigma_t^2}\right)$
- $\prod_{t=1}^{m} \frac{p_t}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(X-\mu_t)^2}{2\sigma_t^2}\right)$

Solution:

$$
\begin{aligned}
p(X = x) &= \sum_{t=1}^{m} p(X = x | T = t) P(T = t) \\
&= \sum_{t=1}^{m} \frac{p_t}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(X-\mu_t)^2}{2\sigma_t^2}\right)
\end{aligned}
$$

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

5. (2)

1/1 point (graded)
Now, given an observation of $X = x$, what is the likelihood that it is drawn from the tth Gaussian distribution?

$$\frac{\frac{1}{\sigma_t}\exp\left(-\left(x-\mu_t\right)^2/\left(2\sigma_t^2\right)\right)}{\sum_{i=1}^m \frac{p_i}{\sigma_i}\exp\left(-\left(x-\mu_i\right)^2/\left(2\sigma_i^2\right)\right)}$$

$$\frac{\frac{p_t}{\sigma_t}\exp\left(-\left(x-\mu_t\right)^2/\left(2\sigma_t^2\right)\right)}{\sum_{i=1}^m \frac{1}{\sigma_i}\exp\left(-\left(x-\mu_i\right)^2/\left(2\sigma_i^2\right)\right)}$$

$$\frac{\frac{p_t}{\sigma_t}\exp\left(-\left(x-\mu_t\right)^2/\left(2\sigma_t^2\right)\right)}{\sum_{i=1}^m \frac{p_i}{\sigma_i}\exp\left(-\left(x-\mu_i\right)^2/\left(2\sigma_i^2\right)\right)}$$

$$\frac{\frac{1}{\sigma_t}\exp\left(-\left(x-\mu_t\right)^2/\left(2\sigma_t^2\right)\right)}{p_t\sum_{i=1}^m \frac{p_i}{\sigma_i}\exp\left(-\left(x-\mu_i\right)^2/\left(2\sigma_i^2\right)\right)}$$

Solution:

Using Bayesian theorem, we have

$$\begin{aligned} P\left(T=t|X=x\right) &= \frac{p\left(X=x|T=t\right)P\left(T=t\right)}{p\left(X=x\right)} \\ &= \frac{\frac{p_t}{\sqrt{2\pi\sigma_t^2}}\exp\left(-\left(X-\mu_t\right)^2/2\sigma_t^2\right)}{\sum_{i=1}^m \frac{p_i}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\left(X-\mu_i\right)^2/2\sigma_i^2\right)} \\ &= \frac{\frac{p_t}{\sigma_t}\exp\left(-\left(x-\mu_t\right)^2/\left(2\sigma_t^2\right)\right)}{\sum_{i=1}^m \frac{p_i}{\sigma_i}\exp\left(-\left(x-\mu_t\right)^2/\left(2\sigma_i^2\right)\right)} \end{aligned}$$

分母是通过total probability theorem得到

Submit You have used 1 of 3 attempts

[STAFF] Q5.3

question posted 2 days ago by quasar_1

On Sept 9, the statement of Q5 (3) was to minimize E[Y-f(X)]. I found this confusing because the value of E[Y-f*(X)] is the same for options (1) and (3) given in the problem (since E[epsilon]=0). Since only one option had to be chosen, I chose option (3) thinking that it would work in bigger generality (even if E[epsilon] is not 0).

Winston_Dai (Staff)

a day ago

We are considering regrade or ungrade this problem.

$\mathbb{E}\left[Y-f\left(X\right)\right]$ was a typo then updated, it should be the squared loss. If not, then $f^*\left(X\right)=\infty$.

@mrBB You are right that all the x in the choices should be X, since X is a r.v. but x is a value.

Sorry for the confusion.

5. (3)

1/1 point (graded)

The objective of regression is to find an optimal function $f^*:\mathbb{R}\to\mathbb{R}$ that minimizes to loss $\mathbb{E}\left[\left(Y-f\left(X\right)\right)^2\right]$ over all choices of f. Suppose we know the generation process in prior (e.g. all the parameters for the multinomial and Gaussian distributions), which of the following is the explicit form of the solution f^*?

$$f^*\left(X\right)=\frac{\sum_{t=1}^m \frac{p_t}{\sigma_t}\exp\left(-\left(x-\mu_t\right)^2/\left(2\sigma_t^2\right)\right)w_tX}{\sum_{i=1}^m \frac{p_i}{\sigma_i}\exp\left(-\left(x-\mu_i\right)^2/\left(2\sigma_i^2\right)\right)}$$

$$f^*\left(X\right)=\frac{\sum_{t=1}^m \frac{p_t}{\sigma_t}\exp\left(-\left(x-\mu_t\right)^2/\left(2\sigma_t^2\right)\right)}{\sum_{i=1}^m \frac{p_i}{\sigma_i}\exp\left(-\left(x-\mu_i\right)^2/\left(2\sigma_i^2\right)\right)w_iX}$$

$$f^*\left(X\right)=\frac{\sum_{t=1}^m \frac{p_t}{\sigma_t}\exp\left(-\left(x-\mu_t\right)^2/\left(2\sigma_t^2\right)\right)\left(w_tX+\epsilon\right)}{\sum_{i=1}^m \frac{p_i}{\sigma_i}\exp\left(-\left(x-\mu_i\right)^2/\left(2\sigma_i^2\right)\right)}$$

I think the thrid choice is factually wrong. We have $f^*\left(X\right)=E\left[Y|X\right]$. Taking this conditional expectation will eliminate any occurrence of RV ε. In other words, $f^*\left(x\right)=E\left[Y|X=x\right]$ should evaluate to a number when substituting the value of an observation x for X. The third answer option evaluates to an expression in ε when substituting $X=x$, which is a random variable: not what we want/expect.

Note to staff: Didn't even realize it when taking the exam, but I think all (small caps) x's in the answer options should be capital X's. (Or alternatively, change the expressions into $f^*\left(x\right)=\dots$ but then X has to be changed into x. Both x's and X's on the RHS of each expression seems inconsistent. Hope I make sense here.

posted 2 days ago by mrBB (Community TA)

Not entirely sure I understand. Are you asking if $f^*\left(X\right)=E\left[Y|X\right]+\varepsilon$ would also minimize $E\left[\left(Y-f^*\left(X\right)\right)^2\right]$? No, it doesn't: $E\left[\left(Y-E\left[X|Y\right]-\varepsilon\right)^2\right]>E\left[\left(Y-E\left[X|Y\right]\right)^2\right)$ (because $E\left[\varepsilon^2\right]=1>0$).

posted 2 days ago by mrBB (Community TA)

- [ ] $$f^*(X) = \frac{\sum_{t=1}^{m} \frac{p_t}{\sigma_t} \exp\left(-(x-\mu_t)^2/(2\sigma_t^2)\right)}{\sum_{i=1}^{m} \frac{p_i}{\sigma_i} \exp\left(-(x-\mu_i)^2/(2\sigma_i^2)\right)(w_i X + \epsilon)}$$ ✔

✔

Grading Note: We will give credit to everyone for this problem because of an error in the original statement of the problem.

Solution:

mrBB (Community TA)
2 days ago - marked as answer a day ago by **Winston_Dai** (Staff)

SInce we have $Y = X + \varepsilon$ in your example, don't we just have $E[Y|X=x] = E[X|X=x] + E[\varepsilon|X=x] = x + 0 = x$ and therefore $E[Y|X] = X$? Can't see where the factor $\frac{1}{2}$ would come from.

To minimize the loss, we need $f^*(X) = \mathbb{E}[Y|X]$.
As $p(Y|X) = \sum_t p(Y,T|X) = \sum_t p(Y|X,T)\,p(T|X)$
we have:

$$\begin{aligned}
f^*(X) &= \mathbb{E}[Y|X] \\
&= \mathbb{E}_{T|X}[\mathbb{E}[Y|T,X]] \\
&= \sum_t^m P(T=t|X)\,\mathbb{E}[Y|T=t,X] \\
&= \sum_t^m \frac{P(X|T=t)\,P(T=t)}{P(X)} w_t X \\
&= \frac{\sum_{t=1}^{m} \frac{p_t}{\sigma_t} \exp\left(-(x-\mu_t)^2/(2\sigma_t^2)\right) w_t X}{\sum_{i=1}^{m} \frac{p_i}{\sigma_i} \exp\left(-(x-\mu_t)^2/(2\sigma_i^2)\right)}
\end{aligned}$$

Submit

You have used 1 of 3 attempts

Answers are displayed within the problem

5. (4)

2/2 points (graded)

Now suppose we don't know the data generation process, but observe N datapoints (x_n,y_n) for $1 \le n \le N$. This time we would like to fit the function f^*, with the constraint that f^* is a linear function. In another word, we would like to find the optimal parameters a^* and b^* for $f^* = a^* X + b^*$ which minimize the empirical loss

$$\sum_{n=1}^{N} (y_n - (a x_n + b))^2$$

over $a \in \mathbb{R}$, $b \in \mathbb{R}$.

Recall in linear regression, we can derive a closed form solution for a^* and b^* by setting the derivative of the loss function to 0. Try to compute this closed form solution and think of the situation when $N \to \infty$, i.e. when we have infinite number of training examples, what is the value of a^* and b^*?

Hint: When $N \to \infty$, $\bar{x} \to \mathbb{E}[X]$, $\bar{y} \to \mathbb{E}[Y]$, $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \to \mathrm{Cov}(X,Y)$, $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \to \mathrm{Var}(X)$, where $\bar{x}$ represents the mean of the observed x, Cov refers to the covariance and Var refers to the variance.

- [ ] $$a^* = \frac{\mathrm{Cov}(X,Y)}{\mathbb{E}[X]}$$

- [ ] $$a^* = \frac{\mathrm{Var}(X)}{\mathrm{Cov}(X,Y)}$$

- [ ] $$a^* = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}$$

- (selected) $a^* = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$

✔

- (selected) $b^* = \mathbb{E}[Y] - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathbb{E}(X)$
- $b^* = \mathbb{E}[X] - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathbb{E}(Y)$
- $b^* = \mathbb{E}[Y] - \frac{\mathrm{Var}(X)}{\mathrm{Cov}(X,Y)}\mathbb{E}(X)$
- $b^* = \mathbb{E}[X] - \frac{\mathrm{Var}(X)}{\mathrm{Cov}(X,Y)}\mathbb{E}(Y)$

✔

Solution:

Taking the derivative of the loss function to a and b, we have

$$\frac{\partial loss}{\partial a} = -2\sum_{n=1}^{N}(y_n - ax_n - b)\,x_n$$
$$\frac{\partial loss}{\partial b} = -2\sum_{n=1}^{N}(y_n - ax_n - b)$$

Set both of them to 0 and we can solve for a and b as

$$\begin{aligned} a^* &= \frac{\frac{1}{N}\sum_{n=1}^{N} x_n y_n - \left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)\left(\frac{1}{N}\sum_{n=1}^{N} y_n\right)}{\frac{1}{N}\sum_{n=1}^{N} x_n^2 - \left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)^2} \\ &= \frac{\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2} \\ b^* &= \frac{1}{N}\sum_{n=1}^{N} y_n - \frac{a^*}{N}\sum_{n=1}^{N} x_n \end{aligned}$$

When $N \to \infty$, we have:

$$\begin{aligned} a^* &= \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \\ b^* &= \mathbb{E}[Y] - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathbb{E}(X) \end{aligned}$$

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

5. (5)

3/3 points (graded)

Now, let's consider a concrete example when $m = 2$, $p_1 = p_2 = 0.5$, $w_1 = 1$,$w_2 = -1$, $\mu_1 = 2$, $\mu_2 = -2$, and $\sigma_1 = \sigma_2 = 1$, what is the value of $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathbb{E}[XY]$? Enter your solutions below.

$\mathbb{E}[X] =$ 0 ✔ Answer: 0

$\mathbb{E}[Y] =$ 2 ✔ Answer: 2

$\mathbb{E}[XY] =$ 0 ✔ Answer: 0

Solution:

The probability density of X now is:

$$p(X) = \frac{1}{2\sqrt{2\pi}}[\exp(-\frac{(x-2)^2}{2}) + \exp[-\frac{(x+2)^2}{2}]]$$

The expectation of X is therefore:

$$\begin{aligned}\mathbb{E}[X] &= \int XP(X)\,dX \\ &= \int X\frac{1}{2\sqrt{2\pi}}[\exp(-\frac{(X-2)^2}{2}) + \exp[-\frac{(X+2)^2}{2}]]\,dX \\ &= \frac{1}{2}[\int X\frac{1}{\sqrt{2\pi}}\exp(-\frac{(X-2)^2}{2})\,dX + \int X\frac{1}{\sqrt{2\pi}}\exp(-\frac{(X+2)^2}{2})\,dX] \\ &= \frac{1}{2}[2+(-2)] \\ &= 0\end{aligned}$$

The distribution of Y here is:

$$\begin{aligned}p(Y) &= p(Y|T=1)\,P(T=1) + p(Y|T=2)\,P(T=2) \\ &= p_1\mathcal{N}(w_1\mu_1, w_1^2\sigma_1^2+1) + p_2\mathcal{N}(w_2\mu_2, w_2^2\sigma_2^2+1) \\ &= \frac{1}{2}\mathcal{N}(2,2) + \frac{1}{2}\mathcal{N}(2,2) \\ &= \mathcal{N}(2,2)\end{aligned}$$

Therefore, $\mathbb{E}[Y] = 2$

Similarly,

$$\begin{aligned}\mathbb{E}[XY] &= \mathbb{E}_T[\mathbb{E}[XY|T]] \\ &= \frac{1}{2}\mathbb{E}[XY|T=1] + \frac{1}{2}\mathbb{E}[XY|T=2] \\ &= \frac{1}{2}\mathbb{E}[w_1X^2+\epsilon X] + \frac{1}{2}\mathbb{E}[w_2X^2+\epsilon X] \\ &= \frac{1}{2}\mathbb{E}[X^2] + \frac{1}{2}\mathbb{E}[-X^2] \\ &= 0\end{aligned}$$

Submit You have used 2 of 5 attempts

ℹ Answers are displayed within the problem

5. (6)

2/2 points (graded)

Given the knowledge of $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ and $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$, what is the value of a^* and b^* in this concrete example, assuming we have infinite number of training data? Enter your solutions below

$a^* =$ 0 ✔ Answer: 0

$b^* =$ 2 ✔ Answer: 2

Solution:

$$\begin{aligned} a^* &= \frac{\text{Cov}(X,Y)}{\text{Var}(X)} \\ &= \frac{\mathbb{E}[XY] - mathbbE[X]\mathbb{E}[Y]}{\mathbb{E}[X^2] - (\mathbb{E}[X])^2} \\ &= 0 \\ b^* &= \mathbb{E}[Y] - a^*\mathbb{E}(X) \\ &= 2 \end{aligned}$$

Submit You have used 2 of 5 attempts

Answers are displayed within the problem

5. (7)

1/1 point (graded)

Does this mean with infinite number of training data, the linear regression model is a good fit for this given scenario? In other words, is the linear regression model a good model for predicting Y from X for $N \to \infty$?

Correction Note (Sept 3): An earlier version does not include the second sentence starting with “in other words".

- ( ) Yes
- (•) No

✔

Solution:

The linear regression gives us the solution $f(X) = a^*X + b^* = 2$ in this concrete example with infinite training data. Apparently this is not a good model to predict Y from X.

Submit You have used 3 of 3 attempts

Answers are displayed within the problem

Error and Bug Reports/Technical Issues

Show Discussion

Topic: Final exam (1 week):Final Exam / Problem 5