Machine learning

Seminar 6

September 19, 2025

Week 1

5. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Which of the following statements about the sigmoid function are correct?

- ☐ $\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a))$.
- ☐ $\sigma(-a) = -\sigma(a)$.
- ☐ $\frac{\sigma(a)-1}{\sigma(a)} = \sigma(a)(\sigma(a) - 1)$.
- ☐ $\sigma(-a) = 1 - \sigma(a)$.

5. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Which of the following statements about the sigmoid function are correct?

- [X] $\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a))$.
- [ ] $\sigma(-a) = -\sigma(a)$.
- [ ] $\frac{\sigma(a)-1}{\sigma(a)} = \sigma(a)(\sigma(a) - 1)$.
- [X] $\sigma(-a) = 1 - \sigma(a)$.

2. Which of the following expressions are correct, given no independence assumption and for (non-trivial) discrete random variables?

- ☐ $\sum_b P(A|B=b) = 1$.
- ☐ For two values $a_1 \neq a_2$, $P(A = a_1 \text{ or } A = a_2|B) = P(A = a_1|B) + P(A = a_2|B)$.
- ☐ $\sum_a \sum_b P(A = a|B = b)P(B = b) = 1$.
- ☐ $p(x) = \int p(x, y)p(y)dy$.

3. Which of the following expressions are correct, given no independence assumption and for (non-trivial) discrete random variables?

- [ ] $\sum_b P(A|B=b) = 1$.
- [X] For two values $a_1 \neq a_2$, $P(A = a_1 \text{ or } A = a_2|B) = P(A = a_1|B) + P(A = a_2|B)$.
- [X] $\sum_a \sum_b P(A = a|B = b)P(B = b) = 1$.
- [ ] None of the above.

4. Which of the following equations are correct?

- ☐ $p(x) = \int p(x|y)dy$.
- ☐ The probability that a continuous random variable x takes on a value on the interval (a, b) with $b > a$ is given by $p(x \in (a, b)) = \int_a^b p(x)dx$.
- ☐ $p(x, y) = p(x|y)p(x)$.
- ☐ None of the above.

4. Which of the following equations are correct?

- [ ] $p(x) = \int p(x|y)dy.$
- [X] The probability that a continuous random variable x takes on a value on the interval (a, b) with $b > a$ is given by $p(x \in (a, b)) = \int_a^b p(x)dx.$
- [ ] $p(x, y) = p(x|y)p(x).$
- [ ] None of the above.

MC: Evil Likelihood Model

1 Suppose you wish to devise an "evil likelihood model", which will make your model likelihood $p(\mathbf{x}|\mathbf{w})$ as low as possible. In this case, we wish to find the minimum likelihood solution for the model parameters $\mathbf{w}$. Which statements are true?

- ☐ In the case of a linear regression, we can find the unique solution to this problem.
- ☐ In the case of linear regression, the minimum likelihood estimate is equal to the maximum likelihood solution, but with opposite signs.
- ☐ The goal of such model is to maximize the function: $-\log p(\mathbf{x}|\mathbf{w})$.
- ☐ The goal of such model is to maximize the function: $\frac{1}{p(\mathbf{x}|\mathbf{w})}$.

1 MC: Evil Likelihood Model

1.0 point · 1 question

Suppose you wish to devise an "evil likelihood model", which will make your model likelihood $p(\mathbf{x}|\mathbf{w})$ as low as possible. In this case, we wish to find the minimum likelihood solution for the model parameters $\mathbf{w}$. Which statements are true?

1.0 point · Multiple choice · 4 alternatives

- [x] The goal of such model is to maximize the function: $\frac{1}{p(\mathbf{x}|\mathbf{w})}$.
- [x] The goal of such model is to maximize the function: $-\log p(\mathbf{x}|\mathbf{w})$.
- [ ] In the case of linear regression, the minimum likelihood estimate is equal to the maximum likelihood solution, but with opposite signs.
- [ ] In the case of a linear regression, we can find the unique solution to this problem.

1 Given a dataset $\mathcal{D} = \{x_n\}_{n=1}^{N}$. The data is normally distributed $\mathcal{N}(x_n|\mu, \sigma^2)$ and we assume a Gaussian prior over $\mu : \mathcal{N}(\mu|0, \sigma_0^2)$. Furthermore, the variance σ^2 is assumed to be known. Let μ_{ML} and μ_{MAP} respectively be the ML and MAP estimates for μ. How does $|\mu_{ML} - \mu_{MAP}|$ change as (i) $\sigma_0 \to 0$, (ii) $\sigma_0 \to \infty$, (iii) $N \to \infty$.

(a) (i) decrease (ii) increase (iii) decrease.

(b) (i) decrease (ii) increase (iii) increase.

(c) (i) increase (ii) decrease (iii) decrease.

(d) (i) increase (ii) decrease (iii) increase.

Given a dataset $\mathcal{D} = \{x_n\}_{n=1}^{N}$. The data is normally distributed $\mathcal{N}(x_n|\mu, \sigma^2)$ and we assume a Gaussian prior over $\mu : \mathcal{N}(\mu|0, \sigma_0^2)$. Furthermore, the variance σ^2 is assumed to be known. Let μ_{ML} and μ_{MAP} respectively be the ML and MAP estimates for μ. How does $|\mu_{ML} - \mu_{MAP}|$ change as (i) $\sigma_0 \to 0$, (ii) $\sigma_0 \to \infty$, (iii) $N \to \infty$.

1.0 point · Multiple choice · 4 choices

- ○ (i) decrease (ii) increase (iii) decrease. 0.0
- ○ (i) decrease (ii) increase (iii) increase. 0.0
- ◉ **(i) increase (ii) decrease (iii) decrease.** 1.0
- ○ (i) increase (ii) decrease (iii) increase. 0.0

2. The Gaussian multivariate distribution for a random variable $\mathbf{x} \in \mathbb{R}^D$ is given by: $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$. Which of the following statements are correct?

- ☐ $\boldsymbol{\Sigma}$ is the covariance matrix of the multivariate Gaussian distribution and is equal to $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$, with $\mathbf{x}$ sampled from the multivariate Gaussian distribution.
- ☐ If $\mathbf{x}$ is a vector of size D (i.e. $\mathbf{x} \in \mathbb{R}^D$), then $\boldsymbol{\mu}$ is also a vector of size D.
- ☐ Let us denote $\mathbf{x} = (x_1, ..., x_D)^T$. If $\boldsymbol{\Sigma}$ is a matrix with nonzero entries on its diagonal, and zero-valued entries for all off-diagonal elements, then $\text{cov}[x_i, x_j] = 0$ for $i \neq j$.

2. The Gaussian multivariate distribution for a random variable $\mathbf{x} \in \mathbb{R}^D$ is given by: $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$. Which of the following statements are correct?

- [ ] $\boldsymbol{\Sigma}$ is the covariance matrix of the multivariate Gaussian distribution and is equal to $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$, with $\mathbf{x}$ sampled from the multivariate Gaussian distribution.
- [X] If $\mathbf{x}$ is a vector of size D (i.e. $\mathbf{x} \in \mathbb{R}^D$), then $\boldsymbol{\mu}$ is also a vector of size D.
- [X] Let us denote $\mathbf{x} = (x_1, ..., x_D)^T$. If $\boldsymbol{\Sigma}$ is a matrix with nonzero entries on its diagonal, and zero-valued entries for all off-diagonal elements, then $\text{cov}[x_i, x_j] = 0$ for $i \neq j$.

Week 2

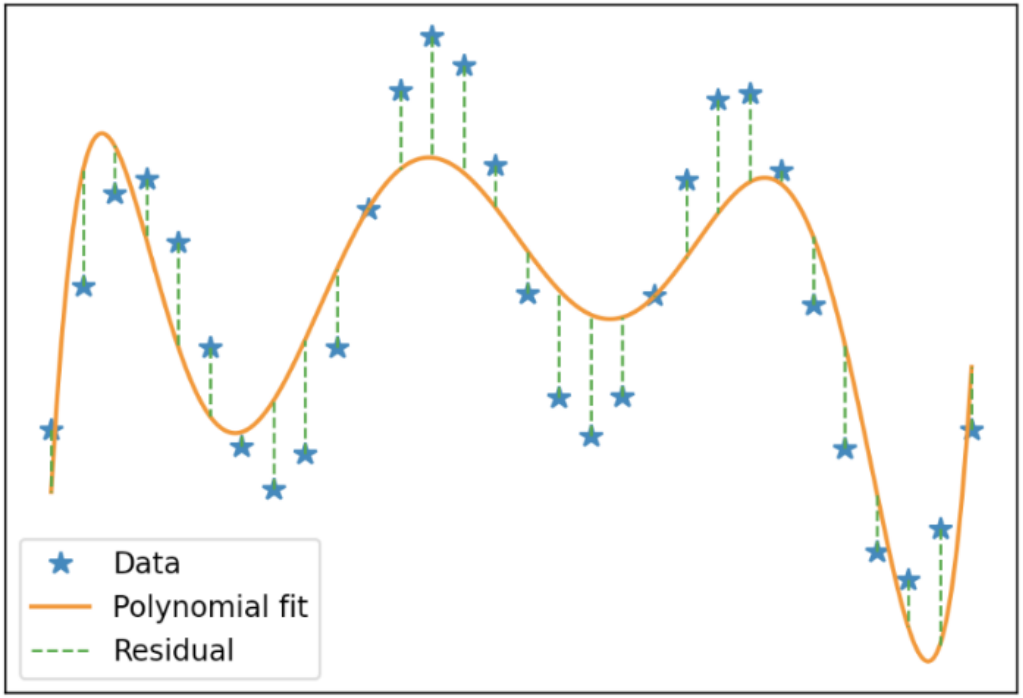
8. Consider two polynomial regression models, model 1 and model 2, of order M_1 and M_2 respecitvely, with $M_1 > M_2$. Which statements are correct?

- [ ] If both models give a low training error, Model M_2 is more likely to lead to a low test error
- [ ] The training error is more likely to be lower for model 2 than for model 1
- [ ] The training error is more likely to be lower for model 1 than for model 2.
- [ ] Model 1 is more sensitive to overfitting than model 2.

8. Consider two polynomial regression models, model 1 and model 2, of order M_1 and M_2 respecitvely, with $M_1 > M_2$. Which statements are correct?

- [X] If both models give a low training error, Model M_2 is more likely to lead to a low test error
- [ ] The training error is more likely to be lower for model 2 than for model 1
- [X] The training error is more likely to be lower for model 1 than for model 2.
- [X] Model 1 is more sensitive to overfitting than model 2.

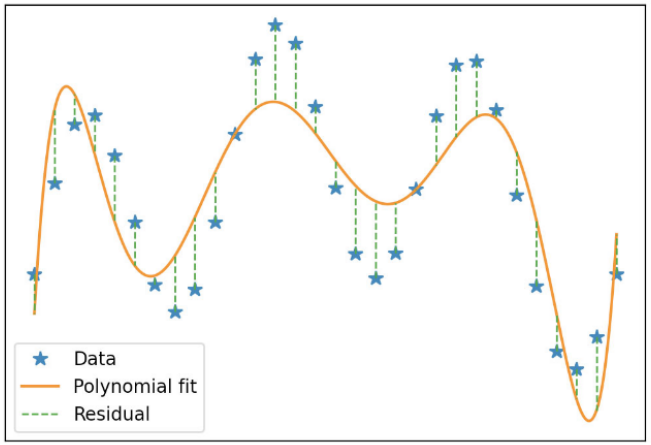
MC: FLAC compression



2

Audio is stored as sequence of measurements of the waveform at discrete time intervals. FLAC is a way of compressing audio by least-squares fitting a polynomial and storing the residual. Weights are stored at full precision, while the residual is compressed further. Which statements are true?

- ☐ The audio data is i.i.d.
- ☐ The weights for the polynomial fit have a closed-form solution.
- ☐ Increasing the order M of the polynomial causes the residual to shrink
- ☐ Adding L2 regularisation can improve compression performance.



Audio is stored as sequence of measurements of the waveform at discrete time intervals. FLAC is a way of compressing audio by least-squares fitting a polynomial and storing the residual. Weights are stored at full precision, while the residual is compressed further. Which statements are true?

1.0 point · Multiple choice · 4 alternatives

- [ ] Adding L2 regularisation can improve compression performance.
- [x] The weights for the polynomial fit have a closed-form solution.
- [ ] The audio data is i.i.d.
- [x] Increasing the order M of the polynomial causes the residual to shrink

MC: Overfitting and model complexity

5 Which of the following statements are true? Check all that apply

- ☐ Higher complexity models are more prone to overfitting and typically have lower variance
- ☐ Only adding more data for training a learner with high bias may not reduce the test error.
- ☐ Overfitting may arise when relevant features are missing in the data
- ☐ Increasing the depth of a neural network will always reduce the test error.

5 MC: Overfitting and model complexity

1.0 point · 1 question

Which of the following statements are true? Check all that apply

1.0 point · Multiple choice · 4 alternatives

- [ ] Higher complexity models are more prone to overfitting and typically have lower variance
- [x] Only adding more data for training a learner with high bias may not reduce the test error.
- [x] Overfitting may arise when relevant features are missing in the data
- [ ] Increasing the depth of a neural network will always reduce the test error.

3. Consider regularized linear regression with the error function $E(\mathbf{w}, \lambda) = \frac{1}{2N} \sum_{n=1}^{N} \left(\boldsymbol{\phi}(\mathbf{x}_n)^T \mathbf{w} - t_n\right)^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$. You want to find the optimal parameter for the regularization penalty $\lambda \in \{0.1,\ 0.01,\ 0.001\}$. You split your dataset consisting of N_{tot} datapoints into a training set, a validation set and a test set. You obtain the following validation and training errors $E_{\text{val}}(\mathbf{w}, \lambda)$, and $E_{\text{train}}(\mathbf{w}, \lambda)$:

	$E_{\text{train}}(\mathbf{w}, \lambda)$	$E_{\text{val}}(\mathbf{w}, \lambda)$
$\lambda_1 = 0.1$	1.72	2.1
$\lambda_2 = 0.01$	0.63	0.85
$\lambda_3 = 0.001$	0.45	1.34

Which of the following statements are correct?

- ☐ If we retrain our model with a larger dataset, we do not need to re-estimate the optimal value of λ.
- ☐ λ_1: overfitting, λ_2: best fit, λ_3: underfitting.
- ☐ λ_1: underfitting, λ_2: best fit, λ_3: overfitting.
- ☐ λ_1: underfitting, λ_2: best fit, λ_3: underfitting.

3. Consider regularized linear regression with the error function $E(\mathbf{w}, \lambda) = \frac{1}{2N} \sum_{n=1}^{N} \left(\boldsymbol{\phi}(\mathbf{x}_n)^T \mathbf{w} - t_n\right)^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$. You want to find the optimal parameter for the regularization penalty $\lambda \in \{0.1,\ 0.01,\ 0.001\}$. You split your dataset consisting of N_{tot} datapoints into a training set, a validation set and a test set. You obtain the following validation and training errors $E_{\text{val}}(\mathbf{w}, \lambda)$, and $E_{\text{train}}(\mathbf{w}, \lambda)$:

	$E_{\text{train}}(\mathbf{w}, \lambda)$	$E_{\text{val}}(\mathbf{w}, \lambda)$
$\lambda_1 = 0.1$	1.72	2.1
$\lambda_2 = 0.01$	0.63	0.85
$\lambda_3 = 0.001$	0.45	1.34

Which of the following statements are correct?

- [ ] If we retrain our model with a larger dataset, we do not need to re-estimate the optimal value of λ.
- [ ] λ_1: overfitting, λ_2: best fit, λ_3: underfitting.
- [X] λ_1: underfitting, λ_2: best fit, λ_3: overfitting.
- [ ] λ_1: underfitting, λ_2: best fit, λ_3: underfitting.

1b Consider building a Bayesian predictive model: let X denote the features, t the label, θ the parameters, $p(\theta)$ the prior, and $p(t|X,\theta)$ the likelihood. Which of the following quantities does the Bayesian framework *ideally* use to make predictions on test data, denoted X^* and t^*?

(a) the posterior mode (a.k.a. MAP estimator): $p(t^*|X^*, \hat{\theta}_{MAP})$

(b) the maximum likelihood estimator (MLE): $p(t^*|X^*, \hat{\theta}_{MLE})$

(c) the posterior predictive distribution: $p(t^*|X^*, t, X) = \int_\theta p(t^*|X^*, \theta) \cdot p(\theta|t, X) d\theta$

(d) the marginal likelihood: $p(t^*|X^*) = \int_\theta p(t^*|X^*, \theta) \cdot p(\theta) d\theta$

1b Consider building a Bayesian predictive model: let X denote the features, t the label, θ the parameters, $p(\theta)$ the prior, and $p(t|X,\theta)$ the likelihood. Which of the following quantities does the Bayesian framework *ideally* use to make predictions on test data, denoted X^* and t^*?

(a) the posterior mode (a.k.a. MAP estimator): $p(t^*|X^*, \hat{\theta}_{MAP})$

(b) the maximum likelihood estimator (MLE): $p(t^*|X^*, \hat{\theta}_{MLE})$

(c) the posterior predictive distribution: $p(t^*|X^*, t, X) = \int_\theta p(t^*|X^*, \theta) \cdot p(\theta|t, X) d\theta$

(d) the marginal likelihood: $p(t^*|X^*) = \int_\theta p(t^*|X^*, \theta) \cdot p(\theta) d\theta$

10. Let $\mathcal{D}$ be the training dataset with an i.i.d. assumption on the data distribution, and $\mathbf{w}$ the model parameters. Which of the following statements about Bayesian linear regression are correct?

- [ ] The standard deviation of the predictive distribution for the target t' of a new datapoint $\mathbf{x}'$ is independent of $\mathbf{x}'$.
- [ ] The predictive distribution for the target t' of a new datapoint $\mathbf{x}'$ takes into account the inherent noise of the true distribution of t', as well as the uncertainty in the values of $\mathbf{w}$.
- [ ] The standard deviation of the predictive distribution decreases when $\mathbf{x}'$ becomes closer to one of inputs $\mathbf{x}_n \in D$.
- [ ] When the number of datapoints in the training dataset D becomes infinite, the standard deviation in the predictive distribution will go to zero.

10. Let $\mathcal{D}$ be the training dataset with an i.i.d. assumption on the data distribution, and $\mathbf{w}$ the model parameters. Which of the following statements about Bayesian linear regression are correct?

- [ ] The standard deviation of the predictive distribution for the target t' of a new datapoint $\mathbf{x}'$ is independent of $\mathbf{x}'$.
- [X] The predictive distribution for the target t' of a new datapoint $\mathbf{x}'$ takes into account the inherent noise of the true distribution of t', as well as the uncertainty in the values of $\mathbf{w}$.
- [X] The standard deviation of the predictive distribution decreases when $\mathbf{x}'$ becomes closer to one of inputs $\mathbf{x}_n \in D$.
- [ ] When the number of datapoints in the training dataset D becomes infinite, the standard deviation in the predictive distribution will go to zero.

13. Let $\mathcal{D}$ be the training dataset with an i.i.d. assumption on the data distribution, and $\mathbf{w}$ the model parameters. Which of the following describes a Bayesian predictive distribution for the target variable t' of a new datapoint $\mathbf{x}'$?

- ☐ $p(t'|D, \mathbf{x}') = \int \int p(t|\mathbf{x}', \mathbf{w}) p(\mathbf{w}) \mathrm{d}\mathbf{w} \mathrm{d}\mathbf{x}'$.
- ☐ $p(t'|D, \mathbf{x}') = \int p(t', \mathbf{w}|\mathcal{D}, \mathbf{x}') \mathrm{d}\mathbf{w}$.
- ☐ $p(t'|D, \mathbf{x}') = \int p(t'|\mathbf{x}', \mathbf{w}) p(\mathbf{w}|D) \mathrm{d}\mathbf{w}$.
- ☐ $p(t'|D, \mathbf{x}') = \int p(t|\mathbf{x}', \mathbf{w}) p(\mathbf{w}) \mathrm{d}\mathbf{w}$.

13. Let $\mathcal{D}$ be the training dataset with an i.i.d. assumption on the data distribution, and $\mathbf{w}$ the model parameters. Which of the following describes a Bayesian predictive distribution for the target variable t' of a new datapoint $\mathbf{x}'$?

- [ ] $p(t'|D, \mathbf{x}') = \int \int p(t|\mathbf{x}', \mathbf{w}) p(\mathbf{w}) \mathrm{d}\mathbf{w} \mathrm{d}\mathbf{x}'$.
- [X] $p(t'|D, \mathbf{x}') = \int p(t', \mathbf{w}|\mathcal{D}, \mathbf{x}') \mathrm{d}\mathbf{w}$.
- [X] $p(t'|D, \mathbf{x}') = \int p(t'|\mathbf{x}', \mathbf{w}) p(\mathbf{w}|D) \mathrm{d}\mathbf{w}$.
- [ ] $p(t'|D, \mathbf{x}') = \int p(t|\mathbf{x}', \mathbf{w}) p(\mathbf{w}) \mathrm{d}\mathbf{w}$.

11. You are given a dataset $\mathcal{D} = \{x_n\}_{n=1}^{N}$. The data is normally distributed $\mathcal{N}(x_n|\mu, \sigma^2)$ and we assume a Gaussian prior over $\mu : \mathcal{N}(\mu|0, \sigma_0^2)$. Furthermore, the variance σ^2 is assumed to be known. Indicate which of the following statements are correct:

- ☐ When N (the number of datapoints in D) becomes larger, the prior distribution will become narrower.
- ☐ When N (the number of datapoints in D) becomes larger, the prior distribution will become wider.
- ☐ When σ_0^2 is small, the maximum a posteriori estimate of μ will be strongly influenced by the prior.
- ☐ When σ_0^2 is large, the maximum a posteriori estimate of μ will be only weakly influenced by the prior.

11. You are given a dataset $\mathcal{D} = \{x_n\}_{n=1}^{N}$. The data is normally distributed $\mathcal{N}(x_n|\mu, \sigma^2)$ and we assume a Gaussian prior over $\mu : \mathcal{N}(\mu|0, \sigma_0^2)$. Furthermore, the variance σ^2 is assumed to be known. Indicate which of the following statements are correct:

- [ ] When N (the number of datapoints in D) becomes larger, the prior distribution will become narrower.
- [ ] When N (the number of datapoints in D) becomes larger, the prior distribution will become wider.
- [X] When σ_0^2 is small, the maximum a posteriori estimate of μ will be strongly influenced by the prior.
- [X] When σ_0^2 is large, the maximum a posteriori estimate of μ will be only weakly influenced by the prior.

Week 3

Probabilistic models

4 In classification, there are three approaches, discriminant functions, probabilistic generative models and probabilistic discriminative models. The following are statements about probabilistic generative models and probabilistic discriminative models. Which of the following statements is true?

- ☐ Logistic regression is a probabilistic discriminative model.
- ☐ In probabilistic discriminative models, the posterior class probabilities $p(C|\mathbf{x})$ are modeled directly.
- ☐ In probabilistic discriminative models, the prior probability of class $p(C)$ is modeled.
- ☐ In generative models, the class conditional probability $p(\mathbf{x}|C)$ are modeled directly.

In classification, there are three approaches, discriminant functions, probabilistic generative models and probabilistic discriminative models. The following are statements about probabilistic generative models and probabilistic discriminative models. Which of the following statements is true?

1.0 point · Multiple choice · 4 choices

☑	**Logistic regression is a probabilistic discriminative model.**	0.34
☑	**In probabilistic discriminative models, the posterior class probabilities $p(C\|\mathbf{x})$ are modeled directly.**	0.34
☐	In probabilistic discriminative models, the prior probability of class $p(C)$ is modeled.	-0.5
☑	**In generative models, the class conditional probability $p(\mathbf{x}\|C)$ are modeled directly.**	0.34

1a Which of the following statements about linear models is **incorrect**?

(a) Linear discriminant analysis and logistic regression have the same expressive power (i.e. in the types of decision boundaries that they can represent).

(b) For logistic regression, the decision boundary lies perpendicular to the parameter vector.

(c) Logistic regression models the difference in the class probabilities, i.e. $P(C_1|X) - P(C_0|X)$, where C_k denotes the class and X the features.

(d) Quadratic discriminant analysis can represent all possible decision boundaries that linear discriminant analysis can represent.

1a Which of the following statements about linear models is **incorrect**?

- ✗ Linear discriminant analysis and logistic regression have the same expressive power (i.e. in the types of decision boundaries that they can represent).
- ✗ For logistic regression, the decision boundary lies perpendicular to the parameter vector.
- (c) Logistic regression models the difference in the class probabilities, i.e. $P(C_1|X) - P(C_0|X)$, where C_k denotes the class and X the features.
- ✗ Quadratic discriminant analysis can represent all possible decision boundaries that linear discriminant analysis can represent.

Misclassification error

3 After fitting a Logistic Regression model with two classes to the training data we calculate the confusion matrix of the model on the test data with N observations: $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$. Which of the following formulas could you use to estimate the misclassification error:

- ☐ $\frac{1}{N}(A + D)$.
- ☐ $1 - \frac{1}{N}(A + D)$.
- ☐ $\frac{1}{N}(B + C)$.
- ☐ $1 - \frac{1}{N}(B + C)$.

After fitting a Logistic Regression model with two classes to the training data we calculate the confusion matrix of the model on the test data with N observations: $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$. Which of the following formulas could you use to estimate the misclassification error:

1.0 point · Multiple choice · 4 choices

- [ ] $\frac{1}{N}(A+D)$. -0.5
- [x] $1-\frac{1}{N}(A+D)$. 0.5
- [x] $\frac{1}{N}(B+C)$. 0.5
- [ ] $1-\frac{1}{N}(B+C)$. -0.5

1 MC: Classification

In the classication setting which of the following statements are true?

- ☐ Probabilistic discriminative models, while estimating $p(C|x)$, often do not need an explicit representation of $p(x|C)$ or $p(x)$.
- ☐ Generative models learn the joint probability distribution $p(x, C)$ and use Bayes' rule to estimate $p(C|x)$
- ☐ Naive Bayes, being a generative model, always outperforms discriminative models when the features are conditionally independent given the class.
- ☐ A perfectly trained logistic regression, as a discriminative model, will always yield the true class posterior probabilities.
- ☐ Generative models inherently allow for a multi-class setting, whereas discriminative models must adopt one-vs-all or one-vs-one schemes.

1 MC: Classification

1.0 point · 1 question

In the classification setting which of the following statements are true?

1.0 point · Multiple choice · 5 alternatives

- [x] Probabilistic discriminative models, while estimating $p(C|x)$, often do not need an explicit representation of $p(x|C)$ or $p(x)$.
- [x] Generative models learn the joint probability distribution $p(x, C)$ and use Bayes' rule to estimate $p(C|x)$
- [ ] Naive Bayes, being a generative model, always outperforms discriminative models when the features are conditionally independent given the class.
- [x] A perfectly trained logistic regression, as a discriminative model, will always yield the true class posterior probabilities.
- [ ] Generative models inherently allow for a multi-class setting, whereas discriminative models must adopt one-vs-all or one-vs-one schemes.

MC: Probabilistic models

7 In classification we consider three models: discriminant functions, probabilistic generative models and probabilistic discriminative models. The following are statements about probabilistic generative models and probabilistic discriminative models. Which are true?

- ☐ Logistic regression is a probabilistic discriminative model.
- ☐ In probabilistic discriminative models, the prior probability of class $p(C)$ is modeled.
- ☐ In generative models, the class conditional probability $p(\mathbf{x}|C)$ are modeled.
- ☐ In probabilistic discriminative models, the posterior probabilities $p(C|\mathbf{x})$ are modeled directly.

7 MC: Probabilistic models

1.0 point · 1 question

In classification we consider three models: discriminant functions, probabilistic generative models and probabilistic discriminative models. The following are statements about probabilistic generative models and probabilistic discriminative models. Which are true?

1.0 point · Multiple choice · 4 alternatives

- [x] Logistic regression is a probabilistic discriminative model.
- [x] In probabilistic discriminative models, the posterior probabilities $p(C|\mathbf{x})$ are modeled directly.
- [ ] In probabilistic discriminative models, the prior probability of class $p(C)$ is modeled.
- [x] In generative models, the class conditional probability $p(\mathbf{x}|C)$ are modeled.